DAVIS, JAMES ROBERT, Ph.D. Exploring Aberrant Responses Using Group Response Functions (2020)
Directed by Dr. Micheline Chalhoub-Deville and Dr. John Willse, 136 pp.

The purpose of this study was to investigate the utility of empirical group response functions (GRFs) for contextualizing aberrant responses in educational test data. A GRF illustrates the functional relationship between expected item proportion correct score (classical *p*-value) and the item difficulty scale, given a specified examinee subgroup and item set or subset. Lack of consistency between empirical and expected (modeled) GRFs may suggest aberrance related to subgroup characteristics, item characteristics, or their interactions. Under relatively ideal simulation conditions, the GRF approach explored in this study appeared to provide an accurate and sensitive representation of subgroup-level aberrance over different item subsets and aberrance types. A demonstration of the approach using real data from a K-12 testing context uncovered a substantive interaction. GRF patterns by item passage type (*informational* versus *literary*) were qualitatively different, but only for examinees designated as English Learners (EL). This result suggests that EL students used different response strategies for different content types, which has implications for both validity and fairness issues. Overall, results provide support that the GRF approach represents a meaningful contribution toward the call for more comprehensive, explanatory person fit methodologies.

EXPLORING ABERRANT RESPONSES USING GROUP RESPONSE FUNCTIONS

by

James Robert Davis

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by

_____
Committee Co-Chair

_____
Committee Co-Chair

To Kara Renea Davis
First the moon. Then the stars.

APPROVAL PAGE

This dissertation written by James Robert Davis has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chair _____

Committee Co-Chair _____

Committee Members _____

_____

_____

March 09, 2020
Date of Acceptance by Committee

March 09, 2020
Date of Final Oral Examination

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge my wife, Kara Davis. There are really no words to express my gratitude for the years of unwavering encouragement, dedication, counsel, and sacrifice. This work is as much her accomplishment as it is mine.

I would like to express my gratitude to the committee co-chairs, Dr. John Willse and Dr. Micheline Chalhoub-Deville. Both individuals have contributed immensely to my professional and academic development. I feel privileged and honored to have had their guidance and support throughout my career as a graduate student.

I would also like to thank committee members Dr. Richard Luecht, Dr. Robert Henson, and Dr. Kinge Mbella. Each of these individuals have provided me with numerous opportunities to develop my research and career interests. I am grateful for the time and expertise they have shared with me, both inside and outside of the classroom.

To all the faculty, staff, and students of the Department of Educational Research Methodology, I could not be more pleased with my decision to join the program. The atmosphere of warmth, collegiality, professionalism, and academic rigor has been integral to my growth. I am especially grateful to Dr. Devdass Sunnassee, Dr. Ayesha Boyce, and Dr. Kyung Yong Kim.

To all my family and friends, my most sincere thank you. I am particularly grateful for the tremendous support of my parents, Jim and Michele Davis, and their continued dedication to their children.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Educational measurement involves making inferences about an examinee's skills, knowledge, or abilities (KSAs) based on responses to tasks, questions, or items on a test. Responses are assumed to be "caused" by the examinee's standing on the KSA or construct of interest (hereafter referred to as a student's level of θ). Given this assumption, θ can then be inferred by item responses (Wilson, 2008). However, when responses are unexpected, the validity of this inference may be threatened. *Person fit* is the extent to which an examinee's response pattern (i.e., responses across multiple items) is consistent with expectation given a particular psychometric model or compared to typical response patterns in a given sample of examinees (Meijer & Sijtsma, 2001). When a response pattern deviates substantially from expectation, such a pattern may be classified as misfitting or *aberrant* (Meijer & Tendeiro, 2014; Petridou & Williams, 2007). An extreme example of aberrance would be a student who gives *correct* responses to the most difficult items while giving *incorrect* responses to the easiest items. This response pattern would be the opposite of expectation and likely not an appropriate indication of the examinee's KSA. Person fit can then be viewed as an evaluation of the extent to which an examinee's estimated θ (given a particular measurement model) predicts his or her item responses (Hambleton et al., 1991). Note that the term aberrance

should not be construed as indicating negative behaviors or deficits for examinees. The term is used throughout only in the strict technical sense: to describe responses that are inconsistent with statistical expectations.

According to The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) in the joint standards (2014; hereafter referred to as the *Standards*), the validity of test score interpretations and uses is a fundamental measurement issue. The *Standards* define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Analysis of response aberrance can be conceptualized as the "final evaluation and confirmation of the validity of the score inference for the proposed use of the test for the individual test taker" (Walker et al., 2018, p. 48). Response patterns that are unexpected given a particular model or context may indicate that factors unrelated to an examinee's standing on the target construct may have influenced responses. The *Standards* refer to such factors as sources of construct-irrelevant variance (CIV).

Researchers have proposed various sources of CIV that may contribute to aberrant responding. Petridou and Williams (2007) summarize these sources and covariates as follows: *demographic characteristics* (e.g., gender, ethnicity, language); *educational characteristics* (e.g., misconceptions, instructional effects); *test-taking strategies* (e.g., guessing, cheating); and *external factors* (e.g., fatigue, noise, distractions). Further, Petridou and Williams (2010) distinguish between "construct-relevant" explanations (e.g., student weaknesses in particular topics) and "construct-irrelevant" explanations

(e.g., item characteristics, such as "wordiness" or item format; test taking strategies, such as "guessing" or "copying"). Exploration of these potential explanations requires comprehensive approaches to person fit analysis (Emons et al., 2005; Mislevy, 2018; Rupp, 2013). Such approaches seek to both identify and contextualize aberrance by combining *global, local,* and *graphical* analyses.

Global analyses use only the information contained within response patterns. Patterns that are substantially inconsistent with a given model or a given sample are flagged as aberrant. Meijer and Sijtsma (2001) provide a methodological review of commonly used global fit indices. The authors review group-based (non-parametric) and IRT-based (parametric) indices. Group-based indices "compare an individual's item-score pattern with the other item-score patterns in the sample" (Meijer & Sijtsma, 2001, p. 109). Many group-based indices are normed against the Guttman (1944) model (e.g., Harnisch & Linn, 1981). "Guttman errors" are defined as an item score pair (0,1) where items in the pair are ordered by difficulty. Such errors count against the fit of a response pattern. Difficulty is defined as the proportion of examinees in the sample that responded correctly (the item *p*-value), which embodies the "group" in such indices.

In contrast to group-based indices, IRT-based indices compare response patterns to what is expected given a psychometric model. Item Response Theory (IRT) represents a family of models that show the relationship between θ and an item response. Various models have been developed for items that are scored dichotomously (e.g., right and wrong). Examples of these models include the Rasch model (Rasch, 1960) and the 2PL and 3PL models (Birnbaum, 1968; Lord & Novick, 1968). Dichotomous IRT models give

the probability of an examinee at a given θ responding correctly to an item. Likelihood-based fit indices (e.g., Drasgow et al., 1985; Snijders, 2001) provide the likelihood of a response pattern given a particular model. Response patterns that are relatively unlikely may be classified as aberrant. Residual-based fit indices (e.g., Wright & Masters, 1982; Wright & Stone, 1979) take into account the differences between observed item scores (e.g., whether a student answered correctly or incorrectly) and expected item scores (e.g., the modeled probability of answering correctly). Differences are aggregated over items, and response patterns with large aggregate differences are classified as aberrant.

After identifying global aberrance, *local* fit analyses seek to contextualize or explain aberrance in terms person characteristics, task or item characteristics, or both. Local analyses take on various forms and conceptualizations. Most often, local analyses explore item-related explanations for aberrance by comparing person fit on different item subsets. For example, a response pattern may be classified as aberrant for the last half of an exam, but not the first half. This result may indicate that the student ran out of time and rapidly guessed on items toward the end of the exam. Various indices and approaches have been developed that explore fit by item subset (e.g., Emons, 2003; Emons et al., 2005; Smith, 1985; Walker et al., 2018). A related area of research is person fit by subgroup (e.g., Cui & Mousavi, 2015; Lamprianou & Boyle, 2004; Meijer et al., 2008; Meijer & Tendeiro, 2014; Meijer & Van Krimpen-Stoop, 2001; Petridou & Williams, 2007). For example, students who speak an additional language at home may be more likely to be classified as aberrant than their mono-linguistic counterparts (Cui & Mousavi, 2015; Petridou & Williams, 2007).

Local fit can also be investigated in terms of the *interactions* between person and task characteristics (e.g., Engelhard et al., 2014). For example, a practitioner may identify subsets of items that examinees engage in differentially based on educational or language-related factors. Mislevy (2018) characterizes local fit analyses as those that incorporate information about individuals and tasks to "explore sociocognitive hypotheses" (p. 249). The sociocognitive testing paradigm describes test performance as arising out of the interaction between *extrapersonal* (social) patterns (e.g., linguistic, cultural, educational) embedded in test tasks and *intrapersonal* (cognitive) patterns that individuals employ to understand and act in the world. From this perspective, the significance of aberrance is not framed in terms of CIV, but in terms of fairness. For example, Mislevy suggests that aberrant patterns "[...] may reflect circumstances of atypical resource development or misunderstanding of the situation" (p. 249). Local fit analyses, then, may also help identify individuals or subgroups who need additional support on specific content.

Statistical analyses (both local and global) can be supplemented with *graphical* procedures. Graphical procedures allow practitioners to communicate aberrance to key stakeholders and further investigate the extremity and nature of aberrant patterns (e.g., Emons et al., 2005; Walker et al., 2018; Walker, Engelhard, Hedgpeth, & Royal, 2016). Common procedures include residual plots (e.g., Ludlow, 1986) and the use of person response functions (e.g., Sijtsma & Meijer, 2001; Trabin & Weiss, 1983). Person response functions (PRFs) are of particular relevance to this dissertation. A PRF models the probability of responding correctly (for a given θ and psychometric model) as a

function of item difficulty. The PRF is monotonic decreasing. As item difficulty increases, probability of a correct response decreases. To illustrate, the PRF for the Rasch model is

$$P\left(X_{kj} = 1 \middle| \theta_k, \delta_j\right) = \frac{1}{1 + \exp\left(\delta_j - \theta_k\right)}, \quad (1)$$

which gives the probability of a fixed person k with θ_k responding correctly to item j ($X_{kj} = 1$) as a function of latent item difficulty, δ. The PRF is analogous to the item response function (IRF), which gives the response probability for a fixed item as a function of latent person proficiency, θ.

The graphical form of the PRF is sometimes referred to as the person characteristic curve (Lumsden, 1977) or person response curve (Trabin & Weiss, 1983). Although these labels are more technically correct, "PRF" is used hereafter to refer to both the function and the graphical representation of the function. Aberrant response patterns can be investigated by comparing the *expected* or modeled PRF [e.g., modeled according to Equation 1] to an *empirical* PRF. Empirical PRFs are generated from the scored response data using one of various techniques (e.g., data smoothing). Deviations of the empirical PRF from the expected PRF are then investigated. Substantial deviations are evidence that the response pattern is not an appropriate indicator of an examinee's θ level. For example, Emons et al. (2004) simulated data and demonstrated deviations of empirical PRFs that may correspond to cheating, anxiety, and guessing. Walker et al. (2018) used empirical PRFs and an interpretive heuristic to assess level of threat posed by

aberrant response patterns to valid score interpretations. The plot in Figure 1.1 gives hypothetical results using Walker et al.'s approach (to be described in detail in Chapter 2). In Figure 1.1(a), the empirical PRF generally follows the contour of the expected PRF with only very small deviations from a monotonically decreasing pattern. Note that each "dot" in the empirical PRFs is an item sorted according to latent item difficulty, δ. This response pattern shows little threat to validity of the score inference. In Figure 1.1(b), the empirical pattern is "W" shaped, which is evidence of serious threat. The hypothetical examinee has several unexpected responses, particularly for the most difficult items. This may be due to, for example, preknowledge of item content.

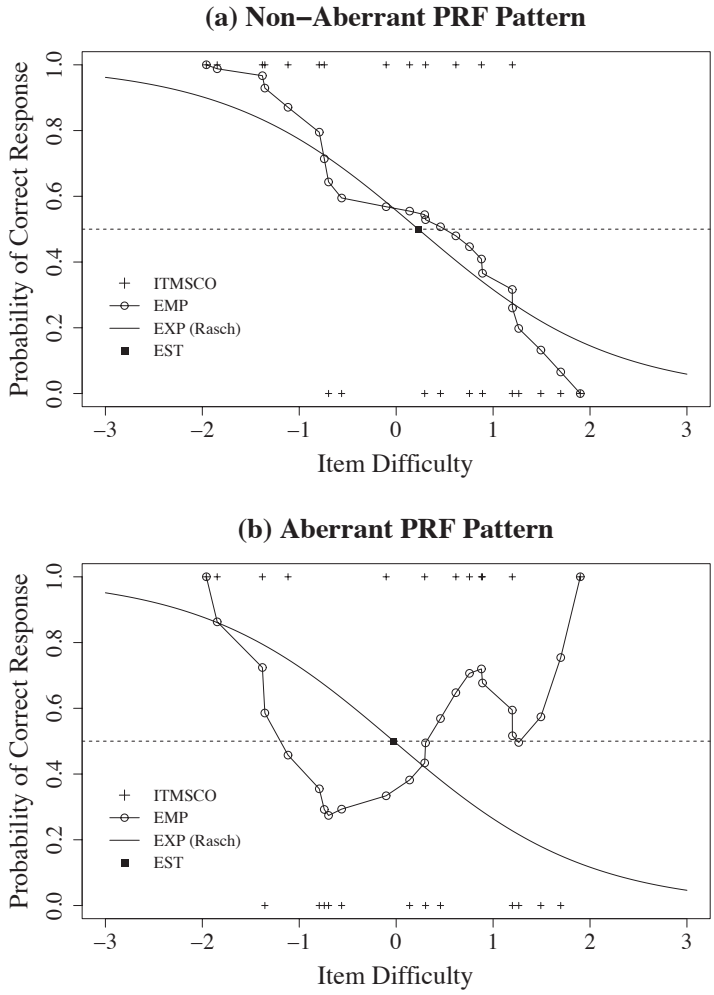


Figure 1.1. Hypothetical Empirical (EMP) and Expected (EXP) PRFs, Dichotomous Item Scores (ITMSCO), and Proficiency Estimate (EST)

Walker et al. (2018) compare PRF patterns for examinees flagged on five fit indices. Two indices were designed to detect global aberrance. And three indices were designed to detect local aberrance related to item characteristics (difficulty, presentation order, and solution strategy). Walker et al. is among the few studies to combine local and graphical analyses. Also, few studies have investigated the interaction between subgroup

and item characteristics to further contextualize aberrant patterns. Moreover, no studies have been found that have explored these interactions graphically, although some researchers have directly called for investigations of PRFs of a group (e.g., Emons et al., 2004). The present study expands on Walker et al.'s approach—and on PRF-based person-fit research more generally—by investigating the utility of *group* response functions (GRFs) for exploring aberrance in terms of interactions between subgroup and item characteristics.

A GRF illustrates the functional relationship between expected *item* proportion correct score (classical *p*-value) and the latent difficulty scale, given a specified examinee group or subgroup. The GRF is analogous to the more familiar *test characteristic curve* (TCC) of item response theory (e.g., Lord & Novick, 1968). The TCC gives the expected proportion correct *for a person* over a set of fixed items (e.g., a test form), whereas the GRF gives the expected proportion correct *for an item* over a group of fixed people. For a fixed examinee group of size K, the expected item proportion correct score, π, given latent item difficulty δ, can be written as

$$\pi(\delta) = \frac{1}{K}\sum_{k=1}^{K} P(X_k = 1|\theta_k, \delta)\,, \tag{2}$$

where $P(X_k = 1|\theta_k, \delta)$ is the PRF [e.g., given by Equation 1].

Lack of consistency between empirical and expected (modeled) GRFs suggests aberrance at the subgroup level. The contour of an empirical GRF, and specifically, where the contour deviates from the expected GRF, may help practitioners identify item

subsets or parts of the scale where subgroups respond most aberrantly. Figure 1.2 illustrates GRF plots for a hypothetical (simulated) subgroup. In Figure 1.2(a), all item responses were generated according to the Rasch model, illustrating a “model fitting”, non-aberrant pattern. The empirical GRF closely aligns with the expected curve. In Figure 1.2(b), the probability of correct response to 12 randomly selected items for 25 randomly selected examinees (50% of total subgroup) was set to 0.95, illustrating a “preknowledge” pattern. In 1.2(c), the probability of correct response to the same 12 items and 25 examinees was set to 0.20, illustrating a “guessing” pattern. In 1.2(d), the guessing data from 1.2(c) was divided into two disjoint item subsets, illustrating a model-fitting pattern on subset 1 (S1) and a guessing pattern on subset 2 (S2), which contained all of the 12 guessing items.

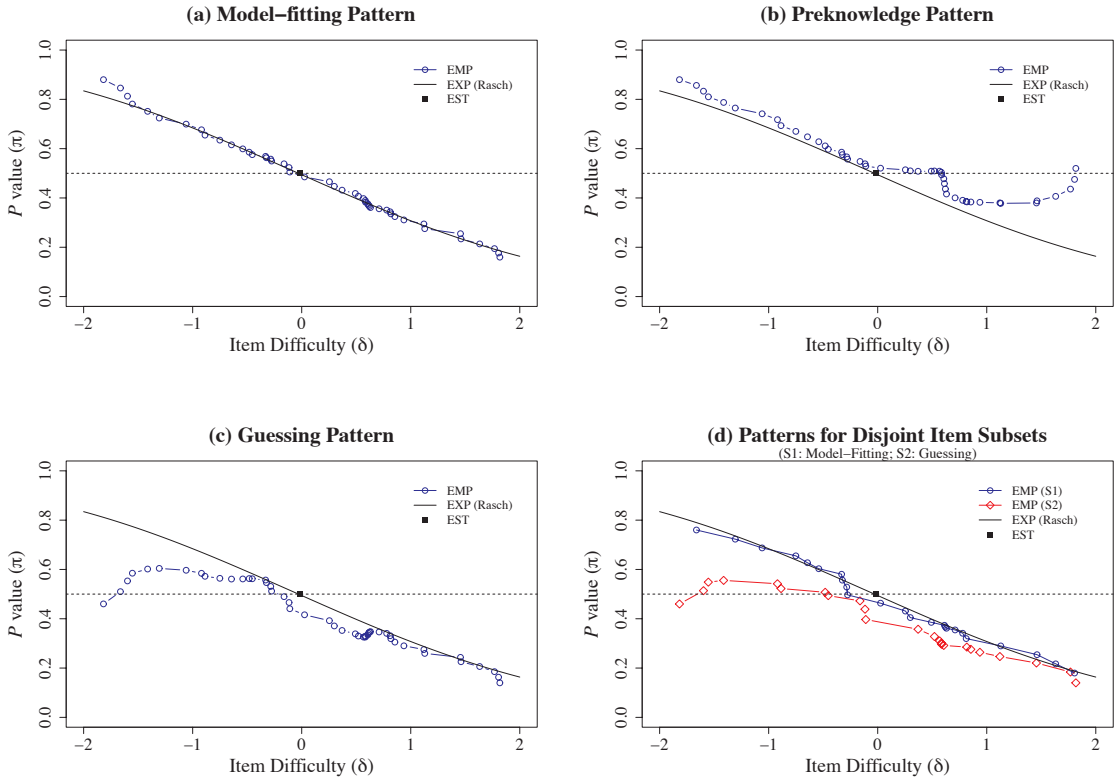


Figure 1.2. Hypothetical Empirical (EMP) and Expected (EXP) GRFs with Subgroup Average Proficiency Estimate (EST; n = 50)

The GRF approach introduced in this study aims to explore and communicate aberrance by comparing patterns *within* subgroup on different item subsets (as in Figure 1.2[d]) and *between* examinee subgroups. Between-subgroup comparisons may be made by comparing plots like Figure 1.2(d) for two or more subgroups. The *accuracy*, *sensitivity*, and *practicality* of the approach will be evaluated in three separate studies, respectively.

Study 1 simulated model-fitting data to investigate whether the smoothing procedure used to produce *empirical* GRFs generates results that conform to the *expected*

GRF. Accuracy was investigated in terms of sampling variance, bias (as defined by Emons et al., 2004), mean absolute deviation (*MAD*), and root mean square error (*RMSE*). Each of these three measures were compared across item subset size, subgroup size, and number of smoothing iterations.

Study 2 provides a simulated demonstration of the sensitivity of the approach to subgroup-based aberrance of various types and severities. Relevant factors are (1) homogeneity of aberrance among subgroup members; (2) number of target aberrant items; and (3) aberrance type. External and internal validity indicators were computed for each condition. The internal indicator is based on the difference in *MAD* values between item subsets. The external indicator is a residual-based, between-fit statistic aggregated across subgroup members. Sensitivity of the GRF to changes in factor levels suggests the approach is useful for identifying differences in the nature and severity of aberrance between item subsets for a particular subgroup. Follow-up analyses investigated the conditions under which the approach may be most sensitive, and therefore most useful or appropriate.

Study 3 provides a real data example. Specifically, the GRF approach was applied to response data from a statewide seventh grade reading test. These data have known differential incidence of aberrance (in terms of global fit) between students designated English Learners (EL) and Non-EL students. Again, the term "aberrance" refers to responses not accounted for by the statistical model. The term is not intended to indicate negative behaviors or any deficits on the part of examinees. To explore local and graphical fit, GRF plots for aberrant EL and aberrant Non-EL subgroups were compared

across subgroups and item subsets. Person fit research using PRFs has provided some guidelines for interpreting inconsistencies between expected and empirical curves (e.g., Emons et al., 2004; Engelhard, 2015; Walker et al., 2018). Some of these guidelines were applied (and adapted) to explore the use of GRFs as a supplement to statistical analyses.

Chapter 2 provides a literature review of studies that have used PRFs to explore person fit. As noted, no literature was found on graphical fit procedures at the group or subgroup level. However, some literature was found on the connection between Differential Item Functioning (DIF) and person fit. Like the GRF approach explored in the present study, DIF analyses attempt to capture differences between subgroups. In turn, sources or causes of DIF may be related to various item characteristics (e.g., linguistic features). The chapter emphasizes the methods and guidelines researchers have used to generate and interpret PRF results. Chapter 3 contains three main sections, each covering one of the three studies mentioned above. Each section provides appropriate detail on procedures, indices, methodological justifications, and anticipated outcomes. Chapter 4 presents results for each of the three studies. Chapter 5 provides a discussion of results and direction for future studies.

CHAPTER II

LITERATURE REVIEW

Fit Analysis using PRFs

No literature was found on group-level response functions for investigating aberrance. Research on PRFs as a tool for person fit analysis is sparse, but a review of this literature will help inform the present study. Trabin and Weiss (1983) framed PRFs as a more general approach than global-level person fit analysis. In global approaches, a single value is obtained that describes the overall fit of a response pattern. In contrast, the PRF describes the entire response pattern as a function of item difficulty, which allows for "the identification of aberrant response patterns or person-item interactions [...] directly" (p. 90). Researchers following Trabin and Weiss frame PRF research as a supplement to global analyses (Emons et al., 2004, 2005; Nering & Meijer, 1998; Sijtsma & Meijer, 2001; Walker et al., 2018; Walker et al., 2016). These studies advocate a more comprehensive approach. For example, Nering and Meijer (1998) suggest "person-fit statistics may be used as a first tool to detect patterns that are unexpected; PRFs can then be inspected to obtain additional information" (p. 54). The following is a brief description of key PRF studies and their contributions to comprehensive approaches to person fit research. In subsequent sections, these studies will be contrasted in terms of theoretical framework, approach to empirical PRF generation, graphical interpretation approaches, and, where applicable, simulation design

Trabin and Weiss (1983) built on previous work that suggested the use of "trace lines" (Weiss, 1973), "subject characteristic curves" (Vale & Weiss, 1975), and "person characteristic curves" (Lumsden, 1977) to describe variability in examinee response behavior. Trabin and Weiss used the term "person response curve" (PRC) to "emphasize that the curve is derived from the responses of an individual to a set of test items" (p. 90). The purpose of their study was to explore the utility of the PRC for investigating unusual variation in examinee responses across items and to investigate the "reliability and other psychometric characteristics" of such a method (p. 90). Using real data, the authors investigated the consistency of empirical PRCs across parallel test forms and the divergence of empirical PRCs from theoretical PRCs given the 3PL model.

Subsequent researchers expanded on Trabin and Weiss (1983) by combining (or comparing) PRF-based person-fit analysis with global fit statistics. The use of PRFs in person-fit research may be statistical, graphical, or both. Statistical approaches provide an index of the divergence of the PRF from the expected curve or, more generally, from a nonincreasing trend (Emons et al., 2004, 2005; Nering & Meijer, 1998; Sijtsma & Meijer, 2001). Such approaches allow quick evaluation of many PRFs, using the PRF essentially as a screening tool to flag aberrant response patterns. Graphical approaches refer to visual inspection of PRF plots. These approaches are more practical after examinees have been flagged, which would reduce the number of plots to inspect. Plots of flagged examinees can be inspected and compared to help practitioners make decisions about the severity and nature of aberrant response patterns (e.g., Walker et al., 2018). For large samples

with a relatively large number of flagged examinees, such an approach may be less practical. A group-level approach may be useful in this situation.

Nering and Meijer (1998) compared the performance of PRFs (using Trabin and Weiss's approach) to the standardized likelihood statistic (l_z; Drasgow et al., 1985). Sijtsma and Meijer (2001) compared a similar PRF approach to Van der Flier's (1982) $ZU3$ global statistic. Both studies simulated data, and both used statistical summaries of the deviation of the empirical PRF from expectation. These summaries are discussed briefly in a subsequent section. In both studies, the non-PRF index generally performed better. However, both studies concluded that PRF information complements global analyses in terms of exploring potential types of misfitting responses (e.g., guessing versus carelessness), as well as identifying potentially problematic items or item subsets.

Emons et al. (2005) proposed a comprehensive approach to person fit analysis, combining both statistical and graphical PRF analyses. They argue that a comprehensive methodology "helps the practitioner to reach a better diagnosis of respondents' misfitting item scores" (p. 102). The approach is in three stages: (1) global analysis using Van der Flier's (1982) *U3* statistic; (2) graphical PRF analysis of response patterns flagged by *U3*; and (3) local analyses to statistically explore empirical PRF deviations from expected trends. The local analysis involved statistically testing the PRF in terms of deviations from a nonincreasing trend for different item subsets (subset by difficulty). For example, the PRF may deviate from the expected trend, but only for high difficulty items, which may indicate answer copying.

Similarly, Walker et al. (2016) framed PRF analysis as supplemental to global analyses, which "may not provide enough detail regarding the nuances of misfit to make a decision regarding the trustworthiness of the score" (p. 41). The authors explored the use of PRFs for identifying different guessing strategies (e.g., no guessing, blind guessing, cued guessing) in a sample of examinees who self-reported their strategy after answering each item. In contrast to Emons et al. (2005), Walker et al. used a graphical (as opposed to statistical) PRF approach only. First examinees were flagged on two global, residual fit indices, *U* and *W* (also known as Mean Square Outfit and Infit, respectively), in a Rasch modeling context (Wright & Masters, 1982; Wright & Stone, 1979). Second, the authors investigated the trend of the empirical PRF against the expected, given by the Rasch model [see Equation 1].

Walker et al. (2018) expanded on Walker et al. (2016) by introducing a heuristic for making decisions about score trustworthiness based on graphical inspection of empirical versus the expected PRF. This heuristic is described in a subsequent section. Walker et al.'s (2018) methodology was to first flag examinees that were aberrant on one or more of five residual fit indices. The authors used two global indices (*U* and *W),* and three local, "between fit" indices (Smith, 1985, 1986), each based on an extension of *U*. The between fit statistic is used to detect differences in fit between two or more items sets. This statistic, as well as *U* and *W,* are described in more detail in Chapter 3. In Walker et al. (2018), the three between fit indices compared fit on item subsets based on difficulty (items either greater than or less than zero logits), presentation order (first versus last half of the exam), and "solution strategy needed to solve the item" (conceptual

or procedural, p. 54). The authors then illustrated PRF patterns that were flagged by different statistics and applied their heuristic.

Theoretical Frameworks for PRFs

As previously noted, a PRF gives the response probability for a person at a specified proficiency level, θ, as a function of item difficulty. Item difficulty may be ordered in terms of the item location parameter defined in an IRT context (e.g., Engelhard, 2015; Trabin & Weiss, 1983; Walker et al., 2018). Item difficulty may also be defined within a non-parametric IRT (NIRT) context as one minus the classical *p*-value (e.g., Emons et al., 2004, 2005; Nering & Meijer, 1998). Regardless of the context, IRT or NIRT, two assumptions are required for the PRF to be a useful tool for exploring aberrance. First, item order should be invariant across examinees. Invariant item order (IIO) means that for any two items, i and j, if for a fixed θ, $P_i(\theta) > P_j(\theta)$, then

$$P_i(\theta) \geq P_j(\theta), \text{ for all } \theta,$$

where $P_i(\theta)$ and $P_j(\theta)$ are the item response functions (IRFs) for items i and j, respectively. In other words, the IRFs for all items are non-intersecting. The IIO assumption provides a straightforward and identical interpretation of the PRF for all examinees. If item order varied across examinees, the same item response pattern may produce aberrance for one examinee but not for another, depending on level of θ. If patterns cannot be compared across examinees with a single item order, the advantage of the PRF as a tool for diagnosing potential causes of aberrance (e.g., specific problem items or item subsets) is reduced.

Second, the PRF is assumed to be monotonic decreasing. For two fixed item difficulty values, δ_a^* and δ_b^* (the star indicating the difficulty value need not be an item location parameter from IRT),

$$P_k(\delta_a^*) \geq P_k(\delta_b^*), \text{ whenever } \delta_a^* < \delta_b^* \text{ for all } k \text{ persons.}$$

Deviations from decreasing, then, are assumed to be indications of person-level aberrance.

Of the parametric IRT family of models, IIO and monotonicity hold for the Rasch model, the 1-parameter logistic model (1PLM) and the 1-parameter normal ogive model (Lord, 1952). In terms of NIRT context, the assumptions hold for the double monotonicity model (Mokken & Lewis, 1982). Walker et al. (2018) preferred the Rasch context to investigate PRFs because the Rasch model is widely used in educational testing. Rasch is therefore a practical choice.

Methods of Constructing Empirical PRFs

Two general methods have been proposed for constructing empirical PRFs. Early researchers used a discrete method (Nering & Meijer, 1998; Sijtsma & Meijer, 2001; Trabin & Weiss, 1983). In the discrete method, items are divided into equal interval strata based on difficulty. For a single examinee, the proportion of correct items within each stratum is computed. A plot is then constructed with strata ordered by difficulty on the *x*-axis and the proportion correct on the *y*-axis.

The second approach is smoothing (Emons et al., 2004, 2005; Engelhard, 2015; Walker et al., 2018; Walker et al., 2016). Smoothing involves weighting each item score

of an item vector (ordered by difficulty) to produce a semi-continuous approximation of an unknown function. Emons et al. (2004) argued that smoothing approaches are preferable to discrete approaches. Smoothing does not require arbitrary decisions about number and size of strata and does not reduce information contained in the response pattern by collapsing items.

Two approaches to smoothing PRFs have been discussed in the literature: kernel smoothing (Emons et al., 2004, 2005) and "Hanning" smoothing (Engelhard, 2015; Walker et al., 2018; Walker et al., 2016). Both approaches order items on difficulty and estimate the PRF at each focal point (i.e., a particular difficulty value) by taking the weighted average of the item scores close to the focal point. In kernel smoothing, weights are defined by a user-specified kernel function and bandwidth. In Hanning smoothing, weights are determined only by the two item score values adjacent to the focal point.

Hanning smoothing (Tukey, 1977; Velleman & Hoaglin, 1981) has been suggested for graphical person fit analysis in the Rasch context (Engelhard, 2015). A smoothed estimate for item *j* (*j =1, ..., J)* is given by

$$h_j = (x_{j-1} + 2x_j + x_{j+1})/4 \quad (3)$$

where x_j is the observed item score. The first and last scores (x_1 and x_J) are left unsmoothed. The smoothed score, h_j, replaces x_2 through x_{J-1} . The smoothing algorithm is then repeated on the smoothed x-values (i.e., the h_j values) in an iterative process. Number of iterations controls the amount of smoothing to the item score vector. Engelhard (2015) and Walker et al. (2018) suggest setting number of iterations equal to

the person's number correct score. For example, if a person scored 12 out of 20 items, the Hanning sequence would be iterated 12 times for this person.

Although many non-parametric smoothing methods are available, as noted, only Hanning and kernel smoothing have been found in connection to person fit. Kernel smoothing estimates are weighted averages with weights defined by a kernel function (Ramsay, 1991; Simonoff, 1996). Examples of kernel functions include Uniform, Quadratic, and Gaussian. Other common smoothing methods, not found in the literature as it pertains to person fit, include local polynomial estimators and spline smoothing (Simonoff, 1996). For example, splines are piecewise polynomial functions that attempt to fit a missing expected function to data. For the present study, Hanning was considered because of its relative computational simplicity and prior use in Rasch-based person fit research.

Graphical Interpretation of Aberrant PRFs

Researchers have offered various guidelines for interpreting the severity and nature of aberrance in PRFs. *Severity* refers to the extent of impact to valid score interpretations. *Nature* refers to the potential causes and correlates of aberrance. Walker et al. (2018) suggest that the most common "sources of person misfit" are random guessing and "slipping" (p. 47). Slipping (sometimes referred to as "carelessness") is defined as answering easy items incorrectly and difficult items correctly. Other examples are sleeping (problems getting started), cheating, and plodding (working too slowly and methodically; Meijer & Sijtsma, 2001). Karabatsos (2003) compared detection rates of thirty-six person-fit indices on five simulated behaviors: cheaters, creatives, lucky

guessers, careless, and random guessers. For example, “lucky guessing” was simulated by giving low performers a 25% chance of correct for the most difficult items. Guessing behavior may also involve some degree of knowledge. Walker et al. (2016) explored aberrance in real data for examinees who self-reported their guessing strategy (if they used one) after each item. Guessing strategies included blind (or random) guessing, cued guessing (i.e., option selection based on a cue within the test or item stimulus) and informed guessing (e.g., eliminating incorrect options based on partial knowledge of the subject). The authors reported that while all three strategies produced empirical PRFs that diverged from the expected (under the Rasch model), each produced a slightly different pattern or shape. Also, compared to random guessing, those who reported also using cued and informed guessing appeared to produce better fitting PRFs.

Trabin and Weiss (1983) discussed information that can be derived from the empirical PRF. The abscissa where probability correct is 0.5 (i.e., $P(X = 1) = 0.50$) can be taken as a measure of examinee θ. The steepness of the PRF is associated with measurement precision or the error in measurement for the examinee. Guessing behavior may be inferred from higher than expected probability of correct for the most difficult items. Carelessness may be inferred from lower than expected probability of correct for the least difficult items. Finally, substantial deviations from non-increasing may indicate violations of unidimensionality—i.e., sources of measurement variance beyond θ. Engelhard (2015) refers to “hills” and “valleys” when describing deviations from decreasing. A hill may occur toward the higher end of the difficulty spectrum, which may

indicate guessing. A valley may occur at the lower end of the spectrum, which may indicate carelessness.

Emons et al. (2004) suggest the use of PRFs to display "subsets of item scores that disagree with the expected item scores" and suggest various interpretations of aberrant PRF patterns or shapes (p. 2). Such shapes may be U-shaped, bell-shaped, or horizontal. U-shaped patterns generally indicate aberrance on the most difficulty items and is characterized by the authors as potential cheating behavior (e.g., answer copying). Scores related to such a pattern are "spuriously high." Bell-shaped patterns are caused by aberrance on the least difficult items, particularly for average or higher performing examinees. The authors associate bell-shaped patterns with "spuriously low" scores due to, for example, test anxiety. Horizontal or near-horizontal PRFs are associated with guessing behavior by a low performing or an unmotivated examinee.

Walker et al. (2018) focused on the use of PRFs to determine the severity of the validity threat imposed by aberrant response patterns (as suggested by a statistical fit index). To do this, the authors proposed a heuristic that routes aberrant patterns into three levels of threat: no major threat; some threat; serious threat. No major threat occurs when empirical PRFs are within (or minimally stray from) the conditional standard error of measurement (CSEM) band of the expected PRF (given by the Rasch model in this case). Some threat occurs when a large proportion of the empirical PRF strays from the CSEM band but intersects close to where $P(X = 1) = 0.50$. A serious threat occurs when either the empirical PRF shape is extreme (monotonic increasing, U-shaped, or W-shaped) or

when there is more than one place where the PRF crosses $P(X = 1) = 0.50$. The latter situation would suggest more than one θ estimate is plausible for the examinee.

Investigating Quality of PRF Methods

Several researchers have simulated data to investigate the quality of their proposed PRF approach for detecting different types of aberrant responses. Nering and Meijer (1998) simulated misfit by first forming two groups of simulees: low and high θ groups. Low simulees had $\theta \leq 0$, and high simulees $\theta \geq 0$. For the low group, they generated a "spuriously high" (SH) condition by rescoring selected items as correct. For the high group, they generated a "spuriously low" (SL) condition by giving simulees a 20% chance of correct on selected items. The researchers also manipulated percent of items selected for aberrance. The conditions were 15%, 20%, or 30% of test crossed with three test lengths (55, 121, and 231).

As previously discussed, Nering and Meijer (1998) used a discrete method of generating empirical PRFs. Ordered items were formed into 11 equal-interval strata. For each stratum, two values were computed: one value according to the simulee's observed item responses (i.e., the proportion correct in the stratum); and one value according to the expected proportion correct for the items in the stratum given the simulee's θ and the model. Sensitivity of this method was investigated by a χ^2 goodness-of-fit test between the observed and expected values across the 11 strata. Statistically significant values indicated aberrance. For the SH condition and unconditional on θ, the PRF was equal or better at detecting simulated aberrance than l_z (Drasgow et al., 1985), a widely-used global fit statistic. In most conditions, however, sensitivity was worse for the PRF

method. Nevertheless, as previously noted, the authors suggest that visual inspection of the PRF provides diagnostic information that global fit statistics do not.

Sijtsma and Meijer (2001) simulated data and introduced a new statistic for detecting aberrant PRFs. The statistic, a "cumulative hypergeometric probability" is computed on two item subsets ordered by difficulty. The statistic gives an upper-bound probability of the total score on the easier item subset given the overall total score, number of total items, and number of items on the easier subset. Low probability (e.g., less than .05) is evidence that the PRF increases from subset to subset. Sijtsma and Meijer simulated "carelessness" by making the probability of correct 0.25 of the 5 easiest items on a 40-item test and of the 10 easiest items on an 80-item test. They compared sensitivity of the new statistic to Van der Flier's (1982) $ZU3$ global statistic. Results show that the new method is less sensitive under most conditions. Nevertheless, the authors argue that the new method is useful to confirm or explore aberrance related to particular item subsets (i.e., for local fit analysis).

Emons et al. (2004) provide two simulation studies. The first simulation tested the accuracy of the kernel smoothing method used to generate empirical PRFs. The authors simulated 100 model-fitting response vectors for a 45-item test with item difficulty between 0 and 1 logits and fixed θ of 0.5. Accuracy was estimated using two metrics: (1) sum of squared errors averaged over the 100 replications, which was used as a measure of sampling variance; and (2) "bias." Both metrics compare the empirical and modeled probabilities across PRF focal points (i.e., the points estimated in the smoothing process). The procedure was replicated over three levels of bandwidth to determine which level

produced the best balance between random and systematic error. Emons et al.'s procedure was adapted for use for the present study and will be detailed in Chapter 3.

Emons et al.'s (2004) second simulation investigated the use of logistic regression to approximate the PRF and detect aberrant responses. The study manipulated test length (20 or 40 items), item discrimination ($\alpha = 1$ or $\alpha = 2$), aberrant response behavior (two types), number of misfitting items (5 or 8 for the 20-item test, 5 or 10 for the 40-item test), and θ level (drawn from $N(-1,0.5)$, $N(0,0.5)$, or $N(1,0.5)$). Response behavior was simulated answering copying (fixing high difficulty items to correct) or simulated test anxiety (fixing probability of correct to .25 for easy items). To investigate global misfit, a likelihood ratio test is used to compare the null (0 slope) against the alternative that the slope of the logistic regression is positive. Positive slope represents an increasing trend, which suggests misfit. To investigate local misfit due to cheating or test anxiety—spuriously low scores (bell-shaped trends) or spuriously high scores (U-shaped trends), respectively—a quadratic term was added to the logistic regression. The likelihood ratio test then compared the null (0 slope) against the alternative that the quadratic term is non-zero. Type I error rates were conservative and often zero. Sensitivity or detection rate was generally higher for the cheating condition than the anxiety condition.

Differential Item Functioning and Model-Data Fit

Differential Item Functioning (DIF) refers to when an item performs statistically differently between subgroups after controlling for subgroup differences in proficiency (AERA et al., 2014). The ability of a test to discriminate between examinees with the same proficiency means, necessarily, that examinee performance depends on more than

one dimension (Lord, 1980). As such, the study of DIF, while commonly rooted in the study of bias and fairness in testing (AERA et al., 2014; Cole, 1993; Mislevy, 2018), is fundamentally a study of dimensionality (Dorans & Holland, 1993). Specifically, DIF refers to the dimensionality introduced by the impact of subgroup-related attributes on item performance. Such attributes may be, for example, gender-based (e.g., Baker et al., 2007), linguistic (e.g., Koo et al., 2014), or ethnically related (e.g., Angoff & Ford, 1973; Mitchelson et al., 2009). Many methods of DIF detection have been developed. Common methods include logistic regression (Rogers & Swaminathan, 1993), the Mantel-Haenszel procedure (MH; Holland & Thayer, 1988), and IRT-based methods (e.g., Rasch-Welsh test available in Winsteps®; Linacre, 2016a). An introduction to DIF techniques and DIF-related issues can be found in Clauser and Mazor (1998).

A related phenomenon to DIF is Differential *Person* Functioning (DPF). DPF refers to when a *person* performs statistically differently between item subsets after controlling for subset differences in difficulty (O'Leary & Smith, 2017). DPF analyses are therefore an inversion of DIF analyses. DIF analyses attempt to identify items that favor particular examinee groups, while DPF analyses attempt to identify examinees who "favor" particular item groups (e.g., by item type, operational status, cognitive domain, content domain). DPF may be investigated using many of the same methods common to DIF analysis (e.g., Linacre, 2016b; O'Leary & Smith, 2017; Smith & Davis-Becker, 2011).

Both DPF and DIF can also be conceptualized in terms of model-data fit as violations of the invariance assumption (Engelhard, 2009). The invariance principle in

IRT holds that when model-data fit is close, item parameter estimates are independent of subgroup and examinee proficiency estimates are independent of item subset (Hambleton & Swaminathan, 1985). Engelhard explored this conceptualization in a Rasch-modeling context using data from a grade seven mathematics test. The test was administrated under different conditions (e.g., use of calculator versus standard administration) and taken by different subgroups (students with disabilities [SWDs] and Non-SWD students). DIF, for example, was investigated by modelling the interaction between items, conditions, and subgroups as it pertained to Rasch-based difficulty estimates. Significant interactions indicated that difficulty estimates for the affected items were not invariant over either subgroups or conditions.

Engelhard et al. (2014) explored DPF at the subgroup level (i.e., differential subgroup functioning). Specifically, researchers looked at the interaction between subgroups (categorized by language status and ethnicity) and items (categorized by task type and format) in terms of model-data fit. Fit for examinees who reported that "English was not their best language" (ENBL) was evaluated in relation to a reference scale composed of examinees who reported that English was their best language (EBL). Evidence of substantial subgroup-level misfit indicated potential fairness issues for ENBL examinees. Average subgroup fit was evaluated over different item subsets to further contextualize aberrance.

DIF and DPF have also been used to explore aberrant test responses in connection to potential cheating behavior. O'Leary and Smith (2017) introduced a two-step approach. First, DPF analysis flagged examinees who performed significantly better on

previously exposed operational (i.e., scored) items than on unexposed "security" items, controlling for item difficulty. Flagged examinees were suspected of benefitting from preknowledge of operational item content. DPF results were then used to group examinees into suspects and non-suspects. DIF analysis was then used to flag items that were significantly easier for suspects, controlling for examinee ability. Flagged items were considered potentially compromised.

GRF analysis can be framed within the DIF-DPF context. For example, differences between item subset *expected* GRFs for a given subgroup may suggest DPF. That is, differential subset GRFs may suggest that proficiency estimates (for at least some subgroup members) are not invariant over item subsets. Although not the goal of the present study, exploring the relationships between GRFs, DIF, and DPF represents an area of research that could lead to practical advantages for measurement professionals. For example, GRF analysis could potentially be used as a pre-screening tool before formal DPF analyses. GRFs could also potentially be used to investigate DPF at the aggregate (group) level, analogous to differential test functioning (DTF). DTF occurs when examinees matched on proficiency but from different subgroups have different expected test scores (AERA et al., 2014). By analogy, differential group functioning (DGF) would occur when items matched on difficulty but from different item subsets have different expected *p*-values (because the group or subgroup "functions" differently over the two subsets).

Conclusion

The present study expands on the PRF literature by introducing the GRF as a method of exploring subgroup by item interactions with respect to aberrance. The focus is not on using the GRF as a *statistical* tool to screen aberrance in subgroups, but on the usefulness of the GRF as a *graphical* tool to help diagnose or explain aberrance after it has been detected. Unlike PRF approaches, the GRF approach provides information about subgroup-level response patterns, which may further aid practitioners in explaining aberrance. In addition, the proposed GRF approach may be more practical for investigating large samples of flagged examinees. The present study represents an initial exploration of the accuracy, sensitivity, and practicality of the GRF approach. PRF research explored in this chapter provided guidance in terms of study design and interpretation of results.

Research Questions

Study 1 has one research question with two subparts:

(1) What number of smoothing iterations minimizes the error in estimating GRFs for simulated data?

  (a) How is this number affected by item subset size?

  (b) How is this number affected by examinee subgroup size?

Study 2 has four research questions:

(2.1) To what extend does homogeneity of aberrant responding in the subgroup affect GRF separation between item subsets?

(2.2) To what extent does number of target items affect GRF separation between

item subsets?

(2.3) To what extent does type of aberrance affect GRF separation between item subsets?

(2.4) To what extent does GRF separation correspond to other available measures of person fit (INFIT, OUTFIT, BETWEEN-FIT)?

Study 3 has two research questions. The first question has three subparts.

(3.1) Do GRFs for different item subsets show substantial differences in terms of:

(a) Deviations from monotonic decreasing?

(b) Relationship to the theoretical GRF and average θ estimate for the subgroup? For example, does one subset GRF cross $\pi(\delta) = 0.50$ multiple times?

(c) Pattern type (e.g., U-shaped, bell-shaped, near horizontal), indicating different possible aberrance types (e.g., carelessness, random guessing)?

(3.2) Do findings for question 3.1 differ between subgroups compared on the same item subsets?

CHAPTER III

METHOD

The GRF approach is modeled loosely on Walker et al.'s (2018) PRF approach (described in *Chapter 2*). In general, the GRF approach (1) flags examinees for aberrance; (2) separates aberrant examinees into subgroups in terms of an *a priori* defined classification scheme (e.g., by language status, gender, ethnicity); (3) separates items into two or more subsets in terms of an *a priori* defined classification scheme (e.g., by difficulty, presentation order, content type); and (4) generates GRF plots for each subgroup for each item subset. GRFs can then be compared between subgroups and within subgroups on different item subsets. Study 1 explores *accuracy* of the method used to construct empirical GRFs. Study 2 explores *sensitivity* of the GRF approach to various degrees and types of simulated aberrance. Study 3 explores the *practicality* of the approach using a real data example. As previously noted, these studies represent an initial exploration of the GRF approach. The emphasis here is on providing a proof of concept, using relatively ideal, simulated conditions. Further studies must be conducted to explore the stability of the approach and generalizability across contexts.

Study 1: GRF Accuracy

Smoothing approaches for empirical PRFs attempt to approximate the unknown response function from empirical response data. Hanning smoothing (Tukey, 1977; Velleman & Hoaglin, 1981) has been applied to PRF fit analysis in the Rasch context (e.g., Walker et al., 2018). In the present study, the Hanning function described in Equation (3) was adapted to construct empirical GRFs. First, *p*-values were computed for each item based on subgroup responses. Second, the *p*-values were ordered in terms of latent item difficulty. Third, the Hanning algorithm was applied to the ordered *p*-values.

A smoothed estimate for item *j* (*j =1, ..., J)* was given by

$$h_j = (p_{j-1} + 2p_j + p_{j+1})/4\ , \tag{4}$$

where p_j is the observed item *p*-value computed from subgroup responses to item j. The first and last *p*-values (p_1 and p_J) were left unsmoothed, following Walker et al. (2018). The smoothed *p*-value, h_j, replaces observed *p*-values p_2 through p_{J-1} . The smoothing algorithm was then repeated on the smoothed *p*-values in an iterative process. Appropriate number of iterations is the primary topic of Study 1.

Any smoothing approach involves balancing two types of errors: sampling variance and bias (Ramsay, 1991). In the present context, large sampling variance may result in over-flagging GRF patterns as aberrant (inflated Type I error rate). Large bias, however, may result in "smoothing over" significant deviations from expected patterns (under-flagging). As previously noted, in kernel smoothing approaches (e.g., Emons et al., 2004), the balance between bias and sampling variance is controlled by the user-

specified bandwidth. In Hanning smoothing, the balance can be controlled by manipulating the number of smoothing iterations. Too many iterations may induce bias. Too few may inflate Type I error rate. The recommendation for generating empirical PRFs using Hanning is to set number of iterations equal to the person number-correct score (Engelhard, 2015; Walker et al., 2016; Walker et al., 2018). However, no studies have been found that have systematically investigated the relationship between Hanning iterations and accuracy. Also, for a GRF application, the appropriate number of iterations may be moderated by number of examinees and number of items.

Using simulated data, Study 1 investigated the accuracy of the Hanning procedure for generating empirical GRFs across item subset size, subgroup size, and number of smoothing iterations. Repeated samples of model-fitting (non-aberrant) item response vectors were simulated according to the Rasch model. Person and item parameters used to generate response probabilities were consistent with estimates derived from real test data from a K-12 educational testing context. The observed item difficulty estimates were used (*Mean* = 0.00; *SD* = 0.846; *Min* = -1.54; *Max* = 2.25). Person parameters (θ) were drawn from $N(1,1)$, which was consistent with the observed distribution of person estimates. The study explored 250 conditions obtained by fully crossing three independent variables:

1. **Item subset length (2 levels).** Item subset length had two assigned levels: 12 and 24. The real data test form contains 48 scored items, therefore accuracy for 24 items was compared to accuracy for 12 items. A 24-item subset represents splitting the test into two equal-length subsets, which could represent, for

example, the first and last half of a fixed form exam. A 12-item subset is closer to subset lengths associated with passage-based exams. In the real data test form, each item was associated with one of six passages, with six to nine items per passage.

2. **Subgroup size (5 levels).** This study investigated subgroups of 25, 50, 100, 500, and 1000. The variation in size reflects the fact that while test data for statewide K-12 testing is typically relatively large, some subgroups (e.g., ethnic minorities, English learners) may be relatively small. Recall also that the final subgroup size used in the proposed GRF approach includes only aberrant examinees. The final subgroup size may therefore depend on several factors: (1) the size of the *total* subgroup in the sample; (2) the person fit statistic used; (3) the flagging criteria used; (4) the incidence of aberrance in the subgroup population.
3. **Number of smoothing iterations (25 levels).** One to 24 iterations were explored. For comparison, an unsmoothed *p*-value condition (zero iterations) was also investigated. The zero-iteration condition is worth exploring because for large subgroup sizes, *p*-value may be sufficient to estimate an accurate GRF.

Each condition was replicated 100 times. The number of smoothing iterations is used to answer the research question for Study 1, with factor 1 answering research question 1a and factor 2 answering research question 1b.

Four outcome variables were computed on each of the 250 conditions. First, to investigate sampling variance, the mean (over replications) sum of squared errors was computed. That is, for each condition,

$$SSE_M = \frac{1}{V}\sum_{v=1}^{V}\sum_{j=1}^{J}\left[\hat{\pi}_v(\delta_j) - \pi(\delta_j)\right]^2, \quad (5)$$

where $\hat{\pi}_v(\delta_j)$ is the empirical GRF for replication v ($v = 1, \ldots, V$) at item j with Rasch scale location δ_j. Total squared deviation (*TSD*) was computed as the squared difference between mean empirical GRF (over replications) and theoretical GRF, summed over items. That is,

$$TSD = \sum_{j=1}^{J}\left[\hat{\pi}_v(\delta_j)_M - \pi(\delta_j)\right]^2. \quad (6)$$

Emons et al. (2004) suggest Equations (5) and (6) for investigating empirical PRF sampling variance and "bias", respectively. Note, however, that in Emons et al.'s formulation, SSE_M contains both sampling variance and bias, as it involves differences between true (known) parameters and the estimates of those parameters. In addition, a more meaningful measure than *TSD* is mean absolute deviation (*MAD)*, which puts differences between empirical and theoretical GRFs on the probability scale. For example, a *MAD* value of 0.2 would indicate the average deviation from the theoretical curve (over all items) is 0.2 probability. The calculation for *MAD* is

$$MAD = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\left[\hat{\pi}_v(\delta_j)_M - \pi(\delta_j)\right]^2}. \quad (7)$$

Results of Study 1 were used to guide Study 2 and Study 3 in terms of selecting a number of smoothing iterations that reflects the best compromise between sampling error and bias. To visualize the appropriate number of iterations, for each condition, Root Mean Square Error (*RMSE*) over *j* items for each replication (*v*) was computed and plotted against number of smoothing iterations. Like *SSE, RMSE* includes both sampling variation and bias. Like *MAD*, *RMSE* is on the probability metric. *RMSE* expresses the average error in estimating the GRF. Mean *RMSE* over all replications was computed as

$$RMSE_M = \frac{1}{V}\sum_{v=1}^{V}\sqrt{\frac{1}{J}\sum_{j=1}^{J}\left[\hat{\pi}_v(\delta_j) - \pi(\delta_j)\right]^2}. \quad (8)$$

The "optimal" number of iterations for a particular condition was defined as the iteration number yielding the minimum $RMSE_M$ (or where further iterations yield no further decrease in $RMSE_M$). Although the optimal number was defined in terms of $RMSE_M$, all four measures discussed above were used to interpret results.

Study 2: GRF Sensitivity

In addition to accuracy, the GRF approach should be reasonably sensitive to the severity and nature of aberrance in subgroups. *Severity* refers to the magnitude of aberrance, and *nature* refers to types of aberrance-related behavior (e.g., guessing, cheating) and potential correlates of aberrance (e.g., various item and subgroup

characteristics). As severity and nature of aberrance changes, GRF patterns should reflect—be *sensitive* to—these changes to a reasonable and useful degree. For example, the patterns in Figures 1.2(b) and 1.2(c) are based on simulations of two different types of aberrant behavior—item preknowledge and guessing, respectively. The two patterns are clearly distinguished from each other and from the model-fitting pattern in Figure 1.2(a). Also, the two patterns in Figure 1.2(d) clearly reflect the difference in aberrance severity between the two disjoint item subsets.

Study 2 provides a simulated demonstration of the sensitivity of the GRF approach to three factors hypothesized to influence aberrance severity, nature, or both. These factors are (A, corresponds to research question 2.1) homogeneity of aberrance among subgroup members; (B, corresponds to research question 2.2) number of target aberrant items; and (C, corresponds to research question 2.3) aberrance type. Each condition derived from these factors contained 48 items and 100 simulees. In all conditions, half the items (24) were selected as the *target subset*. The target subset contains the *target items*, which are the items hypothetically related to subgroup status. *Target members* are subgroup members with a higher probability of being aberrant on target items and lower probability on non-target items. Non-target members had the same probability of being aberrant on all 48 items. For any particular member (whether target or non-target), aberrant items (whether target or non-target) are referred to as *affected items.* Affected items were selected using weighted sampling without replacement. For non-target members, all 48 items had uniform weights. Non-target members therefore had the same probability of being aberrant on target and non-target items. For target

members, target items were weighted uniformly so that weights summed to 0.8. That is, there was an 80% chance that an affected item selected for a target member would be a target item. The weights for non-target items for target members summed to 0.2.

Below is a description of the factors listed above and how they were simulated.

A. **Subgroup homogeneity (4 levels; *K**)** refers to the homogeneity of subgroup members with respect to the target items. Complete homogeneity, which is not likely, would mean all members are aberrant on exactly the same items. Complete heterogeneity, also not likely, would mean *no* members are aberrant on the same items. As homogeneity increases, GRF sensitivity should increase. Homogeneity was operationalized by manipulating the number of target members, *K**, (0, 25, 50, and 75 out of 100 members). Zero target members means all members have equal chance of being aberrant on target and non-target items. Seventy-five means that the majority of members have a greater chance of being aberrant on target than non-target items.

B. **Number of target items (4 levels; *J**)** was 0, 6, 9, and 12 out of 24 items in the target subset. GRF sensitivity is expected to increase as *J** increases. Recall that the proposed GRF approach includes only flagged examinees in analyses. To represent this approach, in all conditions, all simulees were simulated to be aberrant. To simplify the simulation design, non-target and target members were simulated to be aberrant on the same number of items in each condition. For example, when *J** = 6, all simulees were aberrant on exactly six items. However,

as previously described, target members had higher probability of being aberrant on target items than on non-target items.

C. **Aberrance type (2 levels)** relates to the strategy or behavior underlying aberrance. Two types were simulated. *Guessing* (GUESS) was simulated as probability of .25 on affected items (chance of randomly guessing the correct option on a four-option multiple choice item). *Spuriously high* (SH) item responses represent item preknowledge or content (or other item features) that are biased in favor of target members. SH was simulated with probability according to the Rasch model, with θ on the unaffected items and $\theta_{SH} = \theta + \phi$ on the affected items. θ_{SH} represents an *increase* in proficiency estimate due to positive bias. ϕ was drawn from $3 * beta(5,5)$, which added an average of 1.5 logits. A *Spuriously low* (SL) condition was not tested in this study, as results from SH would generalize to SL due to the symmetrical nature of the two conditions. SL responding could represent carelessness, anxiety, or item content or features that are biased against target members (i.e., a *decrease* in proficiency).

The influence of homogeneity (Factor A) and number of target items (Factor B) on sensitivity was investigated in separate sub-studies. Each sub-study was in turn repeated over each level of aberrance type (Factor C), for a total of four sub-studies. For homogeneity studies, number of target items was held constant at $J^* = 12$. For number of target item studies, homogeneity was held constant at $K^* = 75$. The four sub-studies are referenced as follows:

- 2A-GUESS: Homogeneity (Factor A) using simulated guessing

- 2A-SH: Homogeneity (Factor A) using simulated spuriously high responding
- 2B-GUESS: Number of target items (Factor B) using simulated guessing
- 2B-SH: Number of target items (Factor B) using simulated spuriously high responding

For each sub-study:

- The goal is to demonstrate *GRF sensitivity over factor levels.* Evidence for sensitivity suggests the approach is useful for identifying differences in the nature and severity of aberrance between item subsets for a given subgroup.
- Plots (described below) and descriptive statistics were generated to investigate the conditions under which the approach may be most sensitive, and therefore most useful or appropriate.
- The target subset (24 items) was randomly selected and the target items were randomly selected from within the target subset. Target and non-target subsets could represent any number of item classification schemes, including item presentation order (first versus last half of test), which may be related to running out of time and rapid guessing on the final test items. Subsets could also represent a variety of item classifications that may be related to subgroup aberrance (e.g., content type, item format, language complexity, information density).
- The same item difficulty parameters were used, which are also the same parameters used in Study 1. Again, these parameters were the observed

difficulty estimates obtained from a calibration of the real data (*Mean* = 0.00; *SD* = 0.846; *Min* = -1.54; *Max* = 2.25).

- The moderating effect of mean subgroup proficiency was also investigated. Three levels of mean proficiency were simulated (Low, Mid, and High mean θ). The three levels, which are described below, were crossed with the levels of the factor under consideration (either K^* or J^*).
- Two types of statistical analyses were provided: *internal* and *external.* Internal evidence was computed directly from the empirical GRFs. External evidence was based on the relationship between the GRFs and established person-fit indices.

Evidence Related to Internal Criteria

Internal evidence included graphical and statistical components. Statistical evidence was based on the difference (or change) in *MAD* values (ΔMAD) between two GRFs—one GRF for each of two item subsets (target and non-target). Using the MAD formula from Study 1 [Equation 7],

$$\Delta MAD = MAD_{Target} - MAD_{NonTarget} \,. \tag{9}$$

For a given subgroup, fifty replications ($V = 50$) were used to obtain the mean empirical GRF for each item subset. The mean GRF for each was then compared to the theoretical GRF to obtain two MAD values, which were then compared. *MAD* was therefore used as an indicator of aberrance at the subgroup level. Significant differences between subsets were regarded as an indicator of aberrance related to item characteristics. Positive ΔMAD

suggests the target subset contains more aberrance (on average, over subset items) than the non-target. For example, Figure 1.2(d) would likely be associated with a moderately high level of positive ΔMAD. If subset S1 also contained substantial aberrance (of the same nature or type), ΔMAD would likely be reduced. Note that the ΔMAD measure has some limitations, but such limitations would be more relevant in real-data analyses. For example, if item subsets were biased in two different directions (e.g., guessing versus cheating), ΔMAD would not be a reliable indicator of the difference. However, for the purposes of Study 2, the index is a reasonable quantitative representation of the GRF plot because only one aberrance type is simulated within any particular condition.

A relative measure of difference in *MAD* values was computed to supplement the *ΔMAD* statistic. The measure was computed as

$$\Delta MAD_R = \frac{\Delta MAD}{\left(MAD_{NonTarget}\right)}, \tag{10}$$

which converts *ΔMAD* into a proportion of *MAD* in the non-target item subset. Holding *ΔMAD* constant, ΔMAD_R increases as the empirical GRF for the non-target subset moves closer to the theoretical GRF. Larger values of ΔMAD_R suggest that aberrance is more confined to the target subset. ΔMAD_R is sensitive not only to differences between empirical and theoretical GRFs (as quantified by *MAD*) but to the differences in the rate of change in *MAD* (from one condition to the next) between target and non-target GRFs. Because the non-target GRF generally has less aberrance (by design), small increases in aberrance from condition to condition may result in relatively larger rates of *MAD* change

(compared to that of the target subset). The difference in rates may, in turn, lead to slightly different conclusions in some of the conditions studied, depending on whether *ΔMAD* or ΔMAD_R is considered.

For each factor level within each sub-study, *ΔMAD* and ΔMAD_R were computed for each of $G = 150$ subgroups. For each level, member k of subgroup g ($g = 1, \ldots, G$) had a proficiency θ_{g_k} generated from $N(\theta_g, .5)$ with θ_g generated from $U(-1,1)$. Three levels of θ_g were generated by classifying each subgroup into one of the following:

LOW: $\theta_g \in [-1, -0.33)$
MID: $\theta_g \in [-0.33, 0.33]$
HIGH: $\theta_g \in (0.33, 1]$

Crossing θ_g classification (3 levels) with the sub-study factor (4 levels for either Factor A or B) produced, for a given sub-study, 12 conditions. Because θ_g was drawn from a uniform distribution, the counts of subgroups in each condition were on average $G_c = 50$.

For each condition, mean *ΔMAD* and mean ΔMAD_R, were computed over all G_c subgroups. For example,

$$\overline{\Delta MAD} = \frac{1}{G_c} \sum_{g=1}^{G} \Delta MAD_g \, . \tag{11}$$

Conditions were then compared. For each sub-study, boxplot comparison charts were produced to evaluate differences between conditions, where each boxplot represents the distribution of ΔMAD values for a particular condition (see example in Figure 3.1). Non-

overlapping distributions indicate a statistically significant difference in mean *ΔMAD* between conditions—and therefore evidence of GRF sensitivity. Boxplot comparisons helped to clarify the conditions under which the GRF approach may be most sensitive.

GRF plots like that of 1.2(d) were generated for subgroups with the minimum and maximum *ΔMAD* values in each condition. To further investigate sensitivity, the "maximum plot" from one condition was compared to the "minimum plot" in the adjacent condition. If boxplots in adjacent conditions have no overlap, the minimum GRF plot is expected to show more separation between subsets than the maximum condition. Figure 3.1 provides an example of such plots.

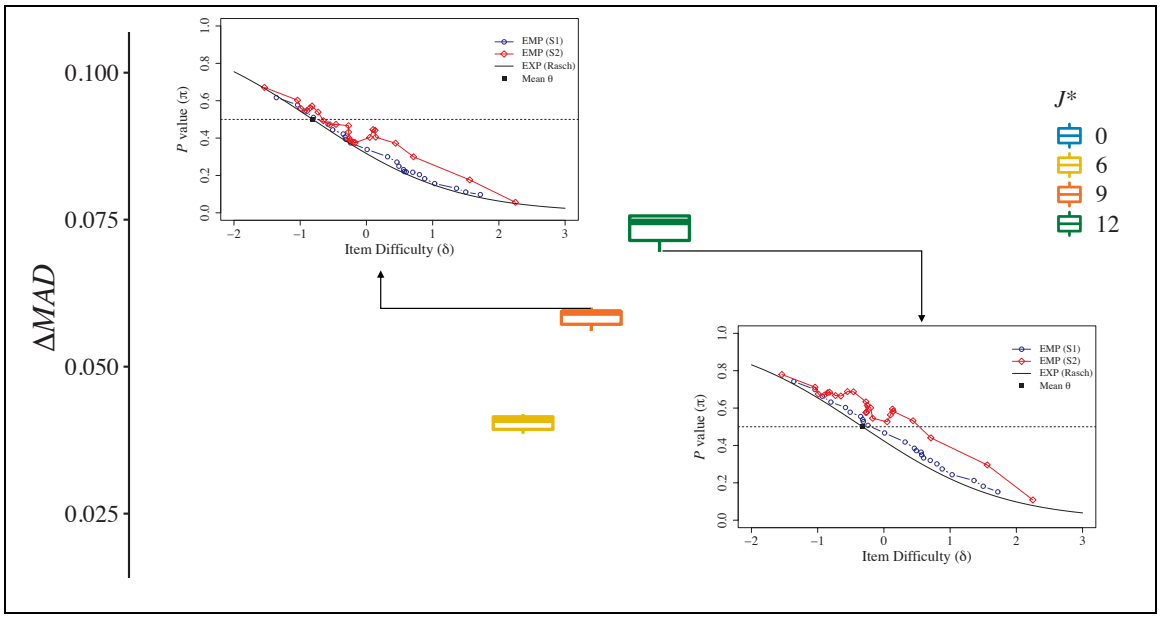


Figure 3.1. Example Boxplot Comparison with Maximum (left) and Minimum (right) GRF Plots for Adjacent Conditions

Evidence Related to External Criteria (Research Question 2.4)

Sensitivity evidence was also derived from external criteria. This part of the study corresponds to research question 2.4: *To what extent does GRF separation correspond to external indices of person fit?* External criteria refer to traditional indices of person fit that are commonly used in a Rasch measurement context. Wright and Stone (1979) and Wright and Masters (1982) developed two residual-based fit statistics, *U* and *W*. Both statistics were developed with the Rasch (1960) model in mind. The unweighted mean squared standardized error for person k is

$$U_k = \frac{1}{J}\sum_{j=1}^{J}\frac{\left(X_{kj} - P_{kj}\right)^2}{P_{kj}(1 - P_{kj})}, \tag{12}$$

where X_{kj} is the dichotomous score, (0,1), for person k on item j, and P_{kj} is the probability of person k correctly answering item j under the Rasch model:

$$P_{kj} = P\left(X_{kj} = 1 \middle| \theta_k, \delta_j\right) = \frac{1}{1 + \exp\left(\delta_j - \theta_k\right)}. \tag{13}$$

U is also known as the Outfit mean-square (Linacre, 2016b), where "Outfit" refers to "outlier sensitive." That is, items that are much more difficult (or much easier) than an examinee is proficient will have a greater impact on the value of *U.* The large difference between difficulty and proficiency is reflected in either very small or very large probability, P_{kj}. In effect, too easy or too difficult items have more influence.

Wright and Masters (1982) proposed the *weighted* mean squared standardized error, *W*, which is less sensitive to item difficulty.

$$W_k = \frac{\sum_{j=1}^{J}(X_{kj} - P_{kj})^2}{\sum_{j=1}^{J} P_{kj}(1 - P_{kj})}. \tag{14}$$

W is also known as the Infit mean-square—Infit referring to "inlier" sensitive (Linacre, 2016). In *W,* the squared residuals are weighted by the variance $P_{kj}(1 - P_{kj})$. When P_{kj} is either very small or very large, variance is small, and the squared residual is penalized to a greater degree.

Both *U* and *W* are centered at one. Values less than one indicate overfit. Values greater than one indicate underfit (or misfit). In addition, both *U* and *W* summarize level of misfit over all J items (i.e., the entire test form). That is, *U* and *W* are global fit indices.

In contrast, Smith (1985) provided a *between-subset* mean square error statistic for the Rasch context. The statistic, UB, was designed to investigate person-level aberrance across two or more disjoint subsets. More specifically, UB can be thought of as the unweighted mean squared error, U, across two or more subsets (Walker et al., 2018). For person k,

$$UB_k = \frac{1}{S-1}\sum_{s=1}^{S}\frac{\left(\sum_{j\in s}^{J_s} X_{kj} - \sum_{j\in s}^{J_s} P_{kj}\right)^2}{\sum_{j\in s}^{J_s}\left[P_{kj} * \left(1 - P_{kj}\right)\right]}, \tag{15}$$

where S is the number of item subsets and J_s is the number of items in subset s. The numerator is the squared difference between observed and expected total scores on subset

s. The expected score, $\sum_{j \in s}^{J_s} P_{kj}$, is the test characteristic function (TCF) evaluated for person k on item subset s. The probability of person k correctly answering item j, P_{kj}, is given by the Rasch model (see Equation 13, above). The denominator of UB is the sum of the variance of P_{kj} over J_s items.

Like U and W, values of UB range from zero to infinity with an expected value of one. Values of UB substantially larger than one suggest "inconsistent performance *between* item subsets (compared to what the model predicts) is present rather than general model inconsistencies over the whole set of test responses" (Walker et al., 2018, p. 54). In other words, UB is designed for more *local*, as opposed to *global,* person fit investigations. Smith (1986) suggested that compared to U, UB is "better at detecting systematic forms of measurement disturbances, e.g., startup, plodding, guessing to complete, and disturbances resulting from specific item/person interactions" (p. 435). Item subsets can be classified according to difficulty, content, presentation order, or any other item characteristic that the researcher suspects may be related to aberrant responding.

For the current study, the mean UB, U, and W for subgroup g was taken over K members and V simulation replications. For example, for UB,

$$UB_M = \frac{1}{K * V} \sum_{v=1}^{V} \sum_{k=1}^{K} UB_{kv} \, . \tag{16}$$

For each sub-study, then, there were $G \times F$ (150 × 4 = 600) simulated subgroups, each with five subgroup-level aberrance indices: ΔMAD, ΔMAD_R, $\overline{UB}$, $\overline{U}$ and $\overline{W}$. The relationship between the internal indices, ΔMAD and ΔMAD_R, with each of the three external indices, was investigated using Spearman's rank correlation coefficient, r_s. Correlation of ranks were used because preliminary analyses strongly suggested a non-linear relationship between ΔMAD and each of the external indices. The correlational analysis was repeated once for each sub-study.

Because all three external indices are measures of fit, ΔMAD (and also ΔMAD_R) is expected to correlate strongly and positively with each external index. However, ΔMAD is expected to correlate most strongly with the between-fit index, $\overline{UB}$, as both indices are measures of local, as opposed to global fit. Large rank correlation between ΔMAD and $\overline{UB}$ would suggest the GRF approach is sensitive to between-subset aberrance with respect to the factor under consideration.

Study 3: Real Data Example

Study 3 provides a real data example of the GRF approach using data from a statewide test of seventh grade English-language arts and reading. The purpose of this example was to explore the practicality of the GRF approach for contextualizing aberrant responses with respect to both subgroup and item subset characteristics. Practicality was addressed in terms of preliminary analyses; applying the GRF procedure; and comparing and interpreting results.

Comparing GRF plots both between subset and between subgroups over different misfit conditions allowed for a graphical investigation of aberrance nature and severity.

Exploration *within* subgroup addressed aberrance related to item characteristics (local fit at the subgroup level). Exploration *between* subgroups addressed potential interactions between subgroup and item characteristics. Specifically, analyses of GRF plots will address the research questions for Study 3, restated here for convenience:

Within Each Subgroup

(3.1) Do GRFs for different item subsets show substantial differences in terms of:

(a) Deviations from monotonic decreasing?

(b) Relationship to the theoretical GRF and average θ estimate for the subgroup? For example, does one subset GRF cross $\pi(\delta) = 0.50$ multiple times?

(c) Pattern type (e.g., U-shaped, bell-shaped, near horizontal), indicating different possible aberrance types (e.g., carelessness, random guessing)?

Between Each Subgroup

(3.2) Do findings for question 3.1 differ between subgroups compared on the same item subsets?

The steps of Study 3 were as follows:

Preliminary Analyses

1. Obtain item and person parameter estimates.
2. Compute U and W for each examinee (global fit).
3. For each examinee, compute UB (between, or local, fit) once for each of three item subset classifications: item difficulty; item order; passage type.

4. For each fit index, apply a bootstrap procedure to compute empirical critical values associated with a 90% confidence interval.
5. Identify misfitting examinees on each index using critical values from (4).
6. Group examinees by language status [English Learners (EL) and Non-EL)]
7. Compare subgroups (statistically and graphically) on global fit distributions, between fit distributions, and proficiency estimate distribution.

GRF Analysis

8. Further classify subgroups by misfit condition, described below. Nine conditions were investigated. Conditions comprise flagging criteria based on measures of person fit.
9. Generate GRF plots by item subset for each subgroup on each misfit condition.
10. Interpret GRF results within and between subgroups. Compare with results from (7).

Data and Instrument

Data came from a test that measures seventh-grade English language arts and reading achievement for students in an Eastern U.S. state. The test contained 48 operational items. All items were multiple choice and scored dichotomously (right or wrong). Items were based on one of six different reading passages with between 6 and 9 items per passage. Reading passages were either informational (e.g., historical, scientific articles) or literary (e.g., short stories, poems). Study 3 was based on data from a single test form containing 53,165 students in the 2016-2017 school year. These data were

chosen because preliminary analysis uncovered statistically significant differential misfit classification between examinees designated as English Learners (EL) and Non-EL students. EL students were classified as aberrant at a higher rate than Non-EL on global fit indices.

Language Learners and K-12 Testing

Although treated as a homogenous group in the context of the present study, the *Standards* (AERA et al., 2014) emphasize that subgroups such as EL students are not homogenous. For example, developmental trajectories of students designated as EL may differ depending on language of instruction. A student who is taught in her native language may have an academic advantage over a student who is taught in English and whose native language learning has been interrupted. However, in English-based testing situations, language would likely be more of a barrier for the former student (i.e., in terms of demonstrating knowledge or skills). Abedi (2005) notes that in part, heterogeneity of language learners (e.g., in terms of culture, literacy, fluency, educational experience, socioeconomic background) makes it difficult to design "a fair assessment system for these students" (p. 180). Similarly, unless specifically accounted for, heterogeneity may also make it difficult to interpret subgroup aberrance. And so, in this study, although in some conditions evidence may suggest no difference between EL a Non-EL subgroups, disaggregating EL examinees (e.g., by socioeconomic status or fluency indicator) may uncover differences.

Preliminary Analyses (Steps 1-7)

Item and person parameter estimates were obtained via Joint Maximum Likelihood Estimation (JMLE) in Winsteps® Rasch measurement software (Linacre, 2016a). JMLE estimates person and item parameters simultaneously by maximizing the likelihood function with respect to both θ, the person proficiency parameter, and δ, the item difficulty parameter (Wright & Panchapakesan, 1969). Assuming local independence holds, the logarithm of the likelihood is given by

$$\ln L(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\delta}) = \sum_{k=1}^{K}\sum_{j=1}^{J}\left[(x_{kj})\ln P_{kj} + \left(1 - x_{kj}\right)\ln(1 - P_{kj})\right], \qquad (17)$$

where x_{kj} is the observed score (0 or 1) for person k on item j and P_{kj} is the expected score under the model. For a Rasch measurement context, the expected score is given by Equation 13. Winsteps® uses successive iterations to adjust estimates until the differences between successive estimates reach a user-specified lower threshold. As these differences approach zero, the likelihood (or log-likelihood) converges to its maximum value (Linacre, 2016b; Wright & Stone, 1979).

Five fit indices were computed for each examinee. Winsteps® provides U and W for each examinee, computed according to Equations (12) and (14), respectively. For each examinee, a between fit value, UB (Equation 15), was computed once for each of three item subset classifications: (1) item difficulty (UB_{diff}); (2) item order (UB_{ord}); and (3) passage type (UB_{pty}). These classifications represent three potential ways practitioners or researchers could investigate local fit. For between fit by item difficulty,

items greater than 0 logits—according to the Rasch difficulty estimate—were in the *difficult* subset (22 items); items less than 0 logits were in the *easy* subset (26 items). For between fit by item order, items were ordered in terms of how they were presented to examinees and split into two equal-length subsets (*first* and *last* half; 24 items in each subset). Finally, for between fit by passage type, items associated with *informational* passages were in one subset (26 items), and items associated with *literary* passages were in the other subset (22 items).

After all indices were computed for each examinee, examinees were then classified as aberrant or non-aberrant (separately, on each index) according to empirically derived critical values. Because Rasch-based fit indices lack known theoretical distributions, this study used a bootstrap procedure to help detect aberrant response patterns. One thousand (1000) replications of model-fitting (non-aberrant) test data were simulated from the Rasch model using item and person parameters taken from calibrations of the real data. For each replication, all fit indices were computed for all simulees. For each index, the 5^{th} and 95^{th} percentile were computed. The mean percentiles for each index were then taken over all replications, giving an estimate of the 90% confidence interval for each index. Point estimates falling outside of the confidence interval were then classified as aberrant. For the present study, only the upper critical value of each index (the 95^{th} percentile) was relevant. Values of a fit index more extreme than the upper critical value represent substantial *inconsistency* with the model or between item subsets. In contrast, values more extreme than the lower critical value represent substantial *consistency.*

Next, examinees were separated into subgroups based on language status (EL, Non-EL). Subgroups were compared in terms of global fit distributions (U and W), between fit distributions (three UBs), and proficiency estimate ($\hat{\theta}$) distribution. Descriptive statistics and density plots were provided for these comparisons. GRF plots were then used to contextualize these analyses.

GRF Analyses (Steps 8-10)

Examinees in each subgroup were further classified into one of six misfit conditions: (1) not flagged on any index (2) flagged on *U*; (3) flagged on *W*; (4) flagged on UB_{ord}; (5) flagged on UB_{pty}; (6) flagged on UB_{diff}. GRF plots were generated for each misfit condition for each subgroup. Based on these conditions, nine (9) GRF analyses were conducted on each subgroup (see Table 3.1). For between fit conditions (conditions 4-6), GRF analyses compared plots by the item subset classification associated with each condition. For example, analyses for subgroups aberrant on UB_{ord} compared GRFs by item order (first and last half of the test).

Table 3.1. GRF Analyses and Associated Misfit Conditions (Study 3)

Analysis	Misfit Condition	Subset Comparison
1	No flags	ORD
2		PTY
3	Flagged on U	ORD
4		PTY
5	Flagged on W	ORD
6		PTY
7	Flagged on UB_{diff}	DIFF
8	Flagged on UB_{ord}	ORD
9	Flagged on UB_{pty}	PTY

Note. U is the outfit statistic, W is the infit statistic, UB_{diff} is between fit by item difficulty, UB_{ord} is between fit by item order, UB_{pty} is between fit by passage type. ORD refers to GRF comparison by item order (first and last half of the test), PTY refers to GRF comparison by passage type (literary versus informational), and DIFF refers to GRF comparison by difficulty (item difficulty estimates either less than or greater than 0 logits).

For each of conditions 1 through 3, two analyses were conducted: comparison of GRF plots by item order (ORD) and by passage type (PTY). Because representation of item difficulty is inherent in a GRF plot, separate analyses of GRFs by item difficulty (DIFF) were not necessary. The first two analyses provided a baseline by involving only examinees who were *not* flagged on any of the statistics. These examinees were grouped into EL and Non-EL, and 10% of each group was randomly sampled. The random sample provided examinee counts that were comparable to the other GRF analyses in Study 3.

CHAPTER IV

RESULTS

Study 1 Results

For each subgroup size, Tables 4.1 and 4.2 provide outcomes (SSE_M, TSD, MAD, and $RMSE_M$) for 12-item subsets and 24-item subsets, respectively. Each table provides results for only a select number of iterations, selected to show, for each condition, the optimal number of iterations. Optimal is defined as the iteration number at which $RMSE_M$ reaches a minimum. The optimal number of iterations for a particular condition is denoted with an asterisk. For each subgroup size, Figures A.1 and A.2 (Appendix A) illustrate the change in $RMSE_M$ over all levels of smoothing iterations (0-24) for 12-item subsets and 24-item subsets, respectively.

Over all conditions, $RMSE_M$ at zero iterations (unsmoothed) had mean 0.04, a minimum of 0.012 (for conditions with 1000-member subgroups), and a maximum of 0.08 (for conditions with 25-member subgroups). The number of optimal smoothing iterations had a mean of 1.4, a maximum of 5 (for 25-member subgroup and 24-item subset), and a minimum of 0 (for conditions with 12-item subsets and subgroup size of 500 or 1000). Decrease in $RMSE_M$ had a mean of 0.015, minimum of 0.001 (for 1000-member subgroup and 24-item subset), and maximum of 0.038 (for 25-member subgroup and 24-item subset). For those conditions where accuracy was improved with smoothing (8 out of 10 conditions), the average decrease in $RMSE_M$ for one iteration was 0.012 and

the average increase in *MAD* was 0.004. As expected, changes in SSE_M over iterations were consistent with $RMSE_M$ findings.

Research Question 1a: Does Item Subset Size Affect Number of Iterations?

Differences in *MAD* values indicate that smoothing tends to have a greater impact on bias for conditions involving 12-item subsets compared to those involving 24-item subsets. For example, the average (over levels of subgroup size) increase in *MAD* over the first three smoothing iterations was 0.017 for 12-item subsets and 0.005 for 24-item subsets. As a result, 1 smoothing iteration (at most) was optimal for 12-item conditions, while up to 5 iterations (at most) was optimal for 24-item conditions. Note however that in any condition, decrease in $RMSE_M$ after the first iteration, if there was a decrease, was negligible.

Research Question 1b: Does Subgroup Size Affect Number of Iterations?

Results suggest that for subgroup sizes greater than or equal to 100, smoothing has either negligible effect on $RMSE_M$ or produces an empirical GRF that is less accurate than the unsmoothed GRF (no smoothing). The largest reduction in $RMSE_M$ was 0.014 (from 0.038 to 0.024) for 100-member subgroup and 24-item subset after 2 smoothing iterations. This can be seen in Figure A.2(b). The increase in *MAD* after 2 iterations was 0.006 (0.005 for *TSD*). For conditions involving subgroups of 500 or 1000, the largest reduction in $RMSE_M$ was 0.004 (from 0.018 to 0.014) for 500-member subgroup and 24-item subset after 1 smoothing iteration. This can be seen if Figure A.2(d). The increase in *MAD* after 1 iteration was 0.004 (0.001 for *TSD*). For conditions with subgroup sizes

greater than 500 and item subset length of 12, any amount of smoothing reduced accuracy. This can be seen in Figures A.1(d) and A.1(e).

Table 4.1. Outcomes for 12 Item Subset (Study 1)

Subgroup Size	Iterations	Outcome			
		***SSE*$_M$**	***TSD***	***MAD***	***RMSE*$_M$**
25	0	0.077	0.001	0.007	0.078
	1*	0.039	0.005	0.010	0.055
	2	0.040	0.012	0.015	0.056
	3	0.045	0.020	0.020	0.059
50	0	0.037	0.001	0.005	0.055
	1*	0.026	0.008	0.013	0.045
	2	0.031	0.016	0.018	0.050
	3	0.039	0.025	0.023	0.056
100	0	0.018	0.000	0.004	0.038
	1*	0.013	0.005	0.010	0.033
	2	0.019	0.012	0.015	0.039
	3	0.026	0.020	0.020	0.046
500	0*	0.004	0.000	0.002	0.018
	1	0.007	0.005	0.009	0.024
	2	0.014	0.012	0.015	0.034
	3	0.022	0.020	0.020	0.042
1000	0*	0.002	0.000	0.001	0.012
	1	0.006	0.005	0.009	0.023
	2	0.013	0.013	0.015	0.033
	3	0.021	0.020	0.020	0.042

Note. SSE_M is the mean (over replications) sum of squared errors; *TSD* is the total squared deviation; *MAD* is the mean absolute deviation; and $RMSE_M$ is the mean (over replications) root mean squared error; * denotes number of iterations at minimum $RMSE_M$.

Table 4.2. Outcomes for 24 Item Subset (Study 1)

Subgroup Size	Iterations	Outcome			
		***SSE*$_M$**	***TSD***	***MAD***	***RMSE*$_M$**
25	0	0.156	0.002	0.007	0.080
	1	0.068	0.002	0.007	0.052
	2	0.054	0.004	0.008	0.046
	3	0.049	0.005	0.009	0.044
	4	0.047	0.007	0.010	0.043
	5*	0.046	0.009	0.011	0.042
	6	0.046	0.011	0.013	0.043
50	0	0.072	0.001	0.004	0.054
	1	0.033	0.002	0.006	0.036
	2*	0.028	0.003	0.008	0.033
	3	0.027	0.005	0.009	0.033
	4	0.027	0.007	0.010	0.033
	5	0.028	0.009	0.012	0.033
100	0	0.036	0.001	0.004	0.038
	1	0.016	0.002	0.005	0.026
	2*	0.014	0.003	0.006	0.024
	3	0.015	0.005	0.008	0.024
500	0	0.008	0.000	0.001	0.018
	1*	0.005	0.001	0.005	0.014
	2	0.005	0.003	0.006	0.015
	3	0.007	0.004	0.008	0.016
1000	0	0.004	0.000	0.001	0.012
	1*	0.003	0.002	0.005	0.011
	2	0.004	0.003	0.006	0.013
	3	0.006	0.005	0.008	0.015

Note. SSE_M is the mean (over replications) sum of squared errors; *TSD* is the total squared deviation; *MAD* is the mean absolute deviation; and $RMSE_M$ is the mean (over replications) root mean squared error. * denotes number of iterations at minimum $RMSE_M$.

For subgroup sizes less than 100 (50 and 25), $RMSE_M$ decrease was *less* negligible than it was for larger subgroup sizes, particularly for the first smoothing iteration. Over both levels of subset size for the first iteration, the average decrease in $RMSE_M$ was 0.02 for subgroups less than 100 and approximately zero for subgroups of 100 or greater. In general, and as expected, there is a greater benefit to smoothing as subgroup size decreases. As subgroup size decreases, the empirical GRF has greater sampling variance and benefits more from smoothing. For larger subgroup sizes, sampling variance is already negligible and any decrease in sampling variance must be weighed against increases in bias.

Study 1 Conclusion

For the conditions studied, error without smoothing was already relatively small, but largest when subgroup sizes were less than 100. The average decrease in $RMSE_M$ due to smoothing was also relatively small, but again largest when subgroup sizes were less than 100. The cost for any reduction in overall error was an increase, albeit small increase, in bias. Increase in bias, as measured by *MAD* and *TSD*, was negligible over the first one to two iterations for all conditions. Real data GRF analyses in statewide K-12 testing contexts may involve subgroup sizes less than 100. As previously noted, final subgroup size will include only those members who were flagged as aberrant, which will depend on fit statistic and flagging criteria. In Study 1, smoothing with at least one iteration had either some beneficial effect or had a negligible effect. A few conditions benefited from more than one iteration. Based on the results, subsequent studies will use the following smoothing guidelines:

1. Subset length approximately 12
    a. Subgroup size 100 or below: one (1) smoothing iteration.
    b. Subgroup size above 100: do not use smoothing.
2. Subset length approximately 24
    a. Subset size 50 or below: three (3) smoothing iterations.
    b. Subset size 51 to 100: two (2) smoothing iterations.
    c. Subset size greater than 100: one (1) smoothing iteration.

The application of these guidelines is expected to provide some improvement to the overall accuracy of empirical GRFs, reducing sampling variance while keeping increase in bias to a minimum.

Study 2 Results

Study 2 was broken up into four sub-studies: subgroup homogeneity using simulated guessing (2A-GUESS); subgroup homogeneity using simulated spuriously high responding (2A-SH); number of target items using simulated guessing (2B-GUESS); and number of target items using simulated spuriously high responding (2B-SH). In each sub-study, 150 subgroups were simulated for each factor level (total of 600 subgroups per sub-study). Each subgroup was randomly assigned a mean proficiency (θ_g) and grouped into a θ_g class (Low, Mid, and High). Crossing the main factor (either homogeneity or number of target items) with θ_g class, there were 12 conditions per sub-study. The mean number of subgroups in each condition across sub-studies was 50 (*SD = 6.67; Min = 38, Max = 64*). Each subgroup contained 100 simulees. Two empirical GRFs of 24 items each were produced for each subgroup. In accordance with the guidelines established in

Study 1, empirical GRFs were smoothed using two (2) iterations. For each sub-study, external and internal criteria were applied to evaluate GRF sensitivity.

Research Questions 2.1–2.3: Results Related to Internal Criteria

These results related to research questions 2.1, 2.2, and 2.3. These results have the potential to interact with each other, so they are presented for each unique combination of conditions (i.e., by sub-study) rather than by research question. However, in the discussion, research questions will be addressed distinctly.

Tables 4.3 and 4.4 provide marginal ΔMAD and ΔMAD_R means for sub-studies involving subgroup homogeneity (Factor A, research question 2.1) and number of target items (Factor B, research question 2.2), respectively. Marginal means are given for (1) levels of the main factor (Factor A or B), averaging over levels of θ_g class and (2) levels of θ_g class, averaging over levels of the main factor. Figures 4.1 through 4.4 provide boxplot comparisons of ΔMAD distributions for each condition for each sub-study. Boxplots that have no overlap between any two conditions are most certainly statistically significant in terms of difference in mean ΔMAD. Boxplots with little to moderate overlap are also likely significant. In the present study, most, if not all, of the boxplots are likely significantly different. The value of the boxplot comparisons, however, is to show trends in GRF sensitivity over factor levels and, in turn, to compare these trends over levels of θ_g class. More boxplot separation indicates greater sensitivity. Appendix B provides GRF plot comparisons of subgroups with the minimum and maximum ΔMAD in adjacent conditions. Adjacent plots were visually inspected for differences in separation

from the theoretical GRF between target and non-target subset empirical GRFs. Appendix B also provides tables of θ_g, *MAD, ΔMAD,* and ΔMAD_R values for the same subgroups.

Sub-study 2A-GUESS. Table 4.3 provides marginal means for subgroup homogeneity for the GUESS condition. Averaging over levels of θ_g class, mean *ΔMAD* increased (ranging from 0.012 to 0.068) as subgroup homogeneity increased. Averaging over levels of subgroup homogeneity, mean *ΔMAD* increased (ranging from 0.022 to 0.055) as θ_g class increased. For each level of θ_g class, boxplots in Figure 4.1 show an increasing trend in *ΔMAD* distribution over levels of subgroup homogeneity. All adjacent boxplots overlapped to some extent in the Low θ_g condition. Non-adjacent boxplots had little to no overlap. For the Mid θ_g condition, boxplots showed more separation than those in the Low θ_g condition. Adjacent conditions $K^* = 0$ and $K^* = 25$ had no overlap, while other adjacent conditions had negligible overlap. All non-adjacent conditions had no overlap. No boxplots overlapped in the High θ_g condition and as a group, had the greater separation than either the Low or Mid θ_g condition. This finding suggests that in the presence of random guessing, the GRF procedure is most sensitive to subgroup homogeneity when subgroups have relatively high mean proficiencies. This finding is expected, as probability of correct, particularly on the higher difficulty items, was already relatively low for lower proficiency simulees.

Table 4.3. Marginal Means for Factor A (*K**) Sub-Studies (Study 2)

Sub-Study	Factor Level	Mean Theta Class	*N* Subgroups	*ΔMAD*		*ΔMAD*$_R$	
				Mean	***SD***	***Mean***	***SD***
2A-GUESS	0	·	150	0.012	0.005	0.201	0.077
	25	·	150	0.032	0.012	0.586	0.099
	50	·	150	0.045	0.020	1.018	0.149
	75	·	150	0.068	0.026	2.158	0.278
	·	Low	215	0.022	0.014	0.885	0.681
	·	Mid	203	0.043	0.023	1.150	0.829
	·	High	182	0.055	0.031	0.938	0.719
2A-SH	0	·	150	-0.011	0.003	-0.156	0.057
	25	·	150	0.014	0.005	0.235	0.070
	50	·	150	0.041	0.009	0.861	0.098
	75	·	150	0.060	0.011	1.644	0.163
	·	Low	188	0.036	0.030	0.796	0.717
	·	Mid	197	0.027	0.027	0.669	0.693
	·	High	215	0.016	0.023	0.493	0.631

Note. *K** is group homogeneity (Factor A); GUESS is simulated guessing; SH is simulated spuriously high responding; *ΔMAD* is the difference in *MAD* values between subset GRFs; *ΔMAD*$_R$ is the ratio of *ΔMAD* to *MAD* for the non-target subset.

GRF plots of minimum and maximum adjacent conditions are only somewhat consistent with the boxplot comparison results. In the Low θ_g condition (Figure B.1), there are only slight differences between adjacent plots, as would be expected given the overlap in boxplots.

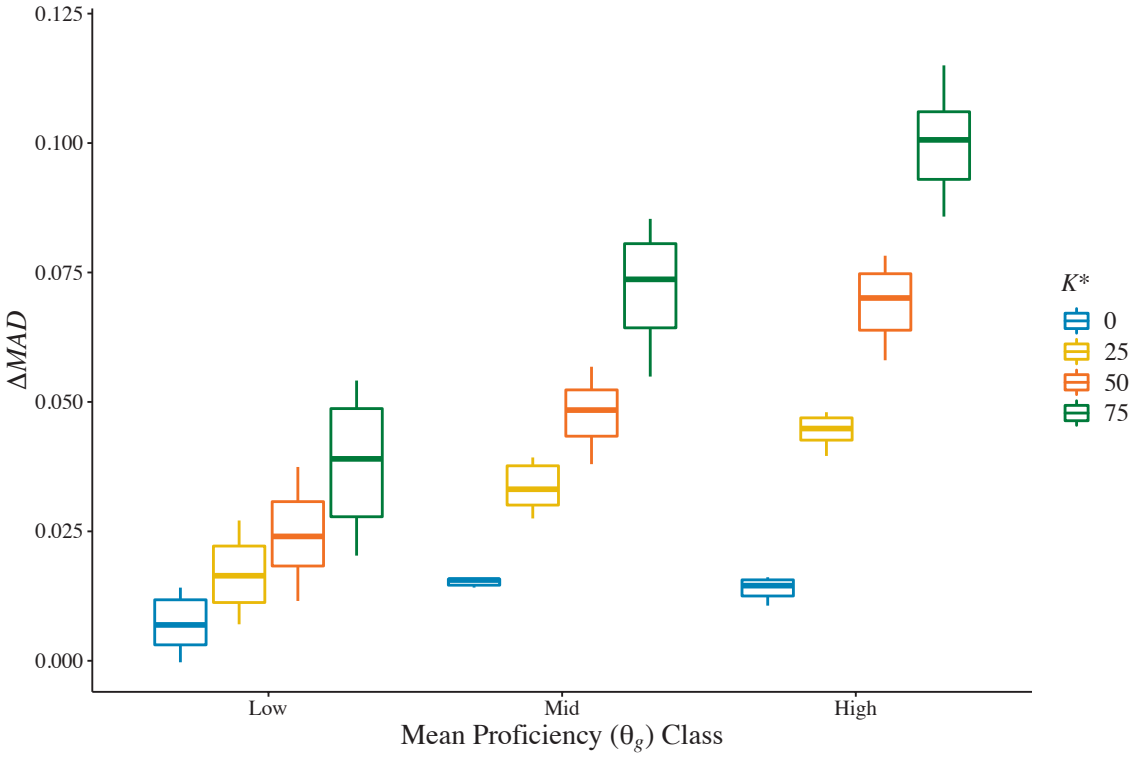


Figure 4.1. Boxplot Comparison of *ΔMAD* by Subgroup Homogeneity (*K**) and Mean Proficiency Class for Guessing (Study 2A-GUESS)

Although boxplots did *not* overlap in the High θ_g condition, there were very subtle visual differences between adjacent GRF plots (Figure B.3). The lack of visual effect may be due to at least two factors: (1) the magnitude of the difference in *ΔMAD* between the plots, and (2) differential rate of decrease in *MAD* between target and non-target subsets, which may also be facilitated by differences in subgroup θ_g. For example, in the High θ_g condition, the maximum subgroup for $K^* = 50$ has $\theta_g = 1.00$ and $\Delta MAD = 0.078$. The minimum plot for $K^* = 75$ has $\theta_g = 0.33$ and $\Delta MAD = 0.086$. The difference in *ΔMAD* is small (0.008), which likely makes visual detection of differences difficult. Also,

because θ_g is smaller for the minimum plot, *MAD* for both the target and non-target subsets decreased relative to *MAD* in the maximum plot. However, the rate of decrease was larger for the non-target subset (approximately 50%) than for the target subset (approximately 20%). The differential rate of decrease had the effect of increasing *ΔMAD* while decreasing the overall aberrance (i.e., $MAD_{Target}+MAD_{NonTarget}$). However, when comparing the *minimum* plots for both conditions, the difference in *ΔMAD* was larger (0.028) and the visual difference was clearer. Differences between other non-adjacent conditions were clear as well (e.g., between $K^* = 25$ *and* $K^* = 75$).

Sub-study 2A-SH. Table 4.3 provides marginal means for subgroup homogeneity for the SH condition. Averaging over levels of θ_g class, mean *ΔMAD* increased as subgroup homogeneity increased (ranging from -0.011 to 0.060). Averaging over levels of subgroup homogeneity, the opposite was true: mean *ΔMAD decreased* as θ_g class increased (ranging from 0.036 to 0.016). This trend is expected, as probability of correct, particularly on the lower difficulty items, was already relatively high for higher proficiency simulees. Artificially inflating proficiency for these simulees had less of an impact than it had for lower proficiency simulees. Boxplot comparisons were consistent with this finding (see Figure 4.2). For each level of θ_g class*, ΔMAD* distribution increased over levels of subgroup homogeneity. For each level of subgroup homogeneity, *ΔMAD* distribution decreased over levels of θ_g class. There was no overlap between any of the boxplots, suggesting that in the presence of spuriously high responding, the GRF

procedure is sensitive to subgroup homogeneity and particularly sensitive for low mean proficiencies.

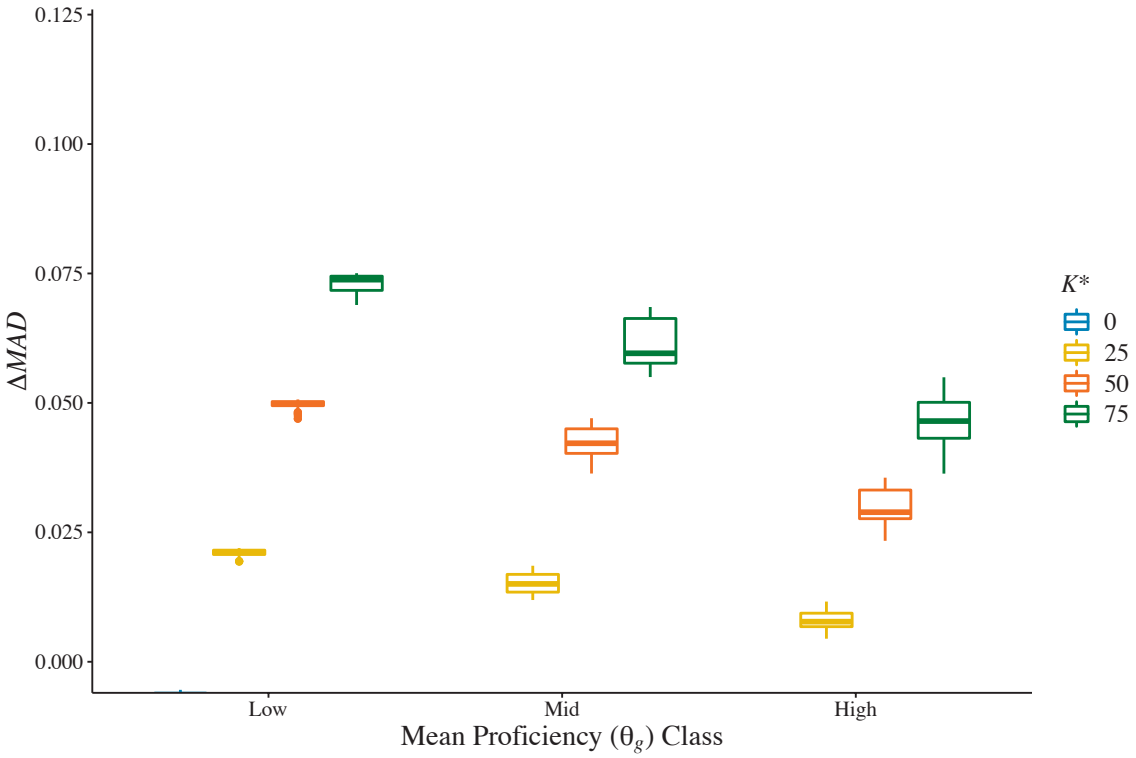


Figure 4.2. Boxplot Comparison of *ΔMAD* by Subgroup Homogeneity (*K**) and Mean Proficiency Class for Spuriously High Responding (Study 2A-SH)

GRF plots of minimum and maximum adjacent conditions are consistent with the boxplot comparison results. Figures B.4, B.5, and B.6 provide adjacent plots (adjacent levels of subgroup homogeneity) for Low, Mid, and High θ_g classes, respectively. In all figures, adjacent plots had clear differences in separation between target and non-target GRFs. Again, the clear visual differences may be due to the magnitude of the difference in mean *ΔMAD* between adjacent conditions. The average difference for this sub-study

was 0.016 (*Min* = 0.001, *Max* = 0.025). Whereas the average difference for the previous sub-study was 0.001 (*Min* = -0.017, *Max* = 0.023).

Table 4.4. Marginal Means for Factor B (*J**) Sub-Studies (Study 2)

Sub-Study	Factor Level	Mean Theta Class	*N* Subgroups	*ΔMAD*		*ΔMAD*$_R$	
				Mean	***SD***	***Mean***	***SD***
2B-GUESS	0	·	150	0.001	0.000	0.138	0.051
	6	·	150	0.038	0.016	2.216	0.413
	9	·	150	0.049	0.022	1.962	0.345
	12	·	150	0.068	0.026	2.158	0.278
	·	Low	185	0.021	0.015	1.357	0.734
	·	Mid	204	0.042	0.027	1.805	0.921
	·	High	211	0.051	0.037	1.667	0.993
2B-SH	0	·	150	0.001	0.000	0.123	0.052
	6	·	150	0.034	0.007	1.857	0.234
	9	·	150	0.050	0.008	1.961	0.108
	12	·	150	0.060	0.011	1.644	0.163
	·	Low	209	0.042	0.027	1.535	0.855
	·	Mid	195	0.039	0.023	1.413	0.710
	·	High	196	0.027	0.018	1.232	0.669

Note. *J** is number of target items (Factor B); GUESS is simulated guessing; SH is simulated spuriously high responding; *ΔMAD* is the difference in *MAD* values between subset GRFs; *ΔMAD*$_R$ is the ratio of *ΔMAD* to *MAD* for the non-target subset.

Sub-study 2B-GUESS. Table 4.4 provides marginal means for Factor B (number of target items) for the GUESS condition. Averaging over levels of θ_g class, mean *ΔMAD* increased (ranging from 0.001 to 0.068) as number of target items increased. Averaging over levels of target items, mean *ΔMAD* increased (ranging from 0.021 to 0.051) as θ_g

class increased. Consistent with sub-study 2A-GUESS, boxplots (Figure 4.3) show a consistent increasing trend over factor levels for each level of θ_g class. Also consistent with 2A-GUESS, separation between boxplots appeared to be greatest for the High θ_g condition. As expected, random guessing behavior is more difficult to detect in low proficiency as opposed to high proficiency subgroups. All adjacent boxplots have some overlap except for the lowest level ($J^* = 0$) with the next highest level ($J^* = 6$). Recall that the lowest level is the non-aberrant condition (i.e., no target items). For non-adjacent conditions there is of course less overlap. This finding suggests that in the presence of random guessing, the GRF procedure is, as would be expected, progressively more sensitive as the difference between number of target items increases (e.g., difference of 3 items versus a difference of 6 items).

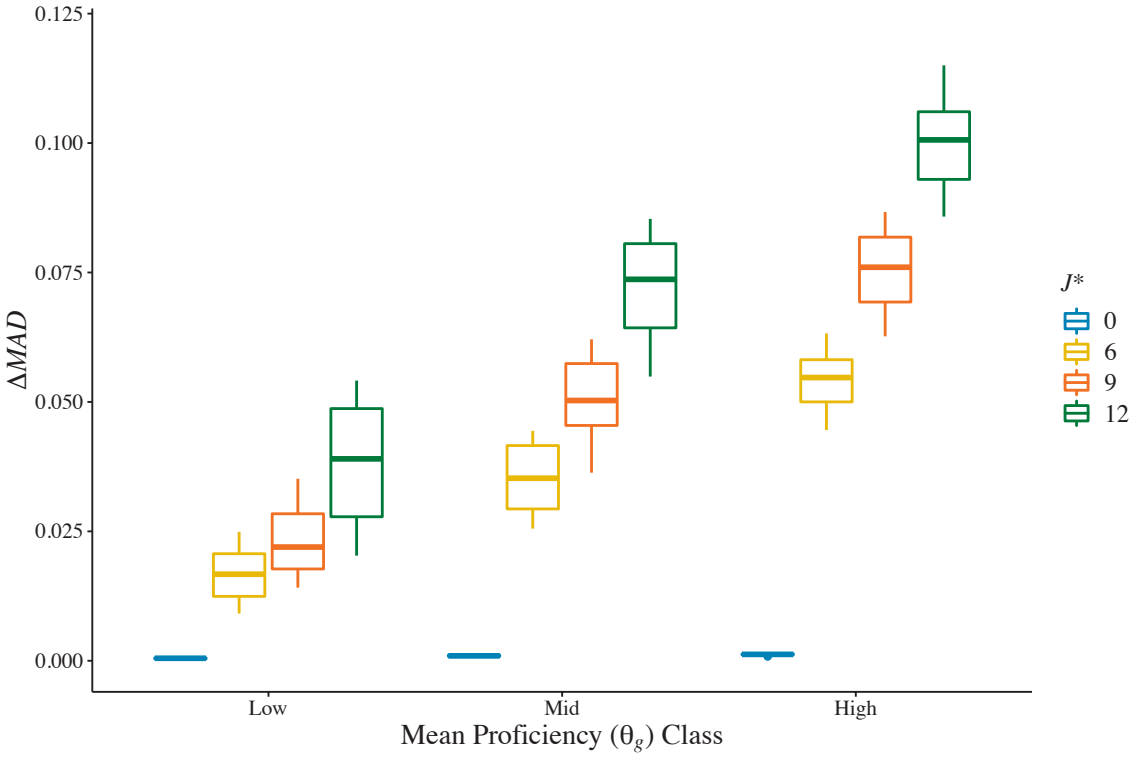


Figure 4.3. Boxplot Comparison of *ΔMAD* by Number of Target Items (*J**) and Mean Proficiency Class for Guessing (Study 2B-GUESS)

GRF plots of minimum and maximum adjacent conditions were consistent with the boxplot comparison results. Over levels of target items, plots had increasing separation between target and non-target GRFs for Low, Mid, and High θ_g classes (Figures B.7, B.8, and B.9, respectively). Between adjacent conditions, the maximum plot tends to have more separation than the minimum plot, which is consistent with the overlap found in the boxplot comparison.

Sub-study 2B-SH. Table 4.4 provides marginal means for Factor B (number of target items) for the SH condition. Averaging over levels of θ_g class, mean *ΔMAD* increased as target items increased (ranging from 0.001 to 0.060). Averaging over levels

of target items, the opposite was true: mean *ΔMAD decreased* as θ_g class increased (ranging from 0.042 to 0.027). This trend is the same as was observed in sub-study 2A-SH. Again, artificially inflating proficiency for high proficiency simulees had relatively little impact on probability of a correct response, particularly for the lower difficulty items. Boxplot comparisons are consistent with this finding (see Figure 4.4). For each level of θ_g class, *ΔMAD* distribution increased over levels of target items (*J**). For each level of *J**, *ΔMAD* distribution decreased over levels of θ_g class. For Low θ_g class, there was no overlap between any of the boxplots. For the Mid and High θ_g classes, there was overlap only between *J** = 9 and *J** = 12. Boxplot results suggest that in the presence of spuriously high responding, the GRF procedure is sensitive to number of target items and particularly sensitive for low mean proficiencies, as the Low θ_g condition appeared to have the most separation.

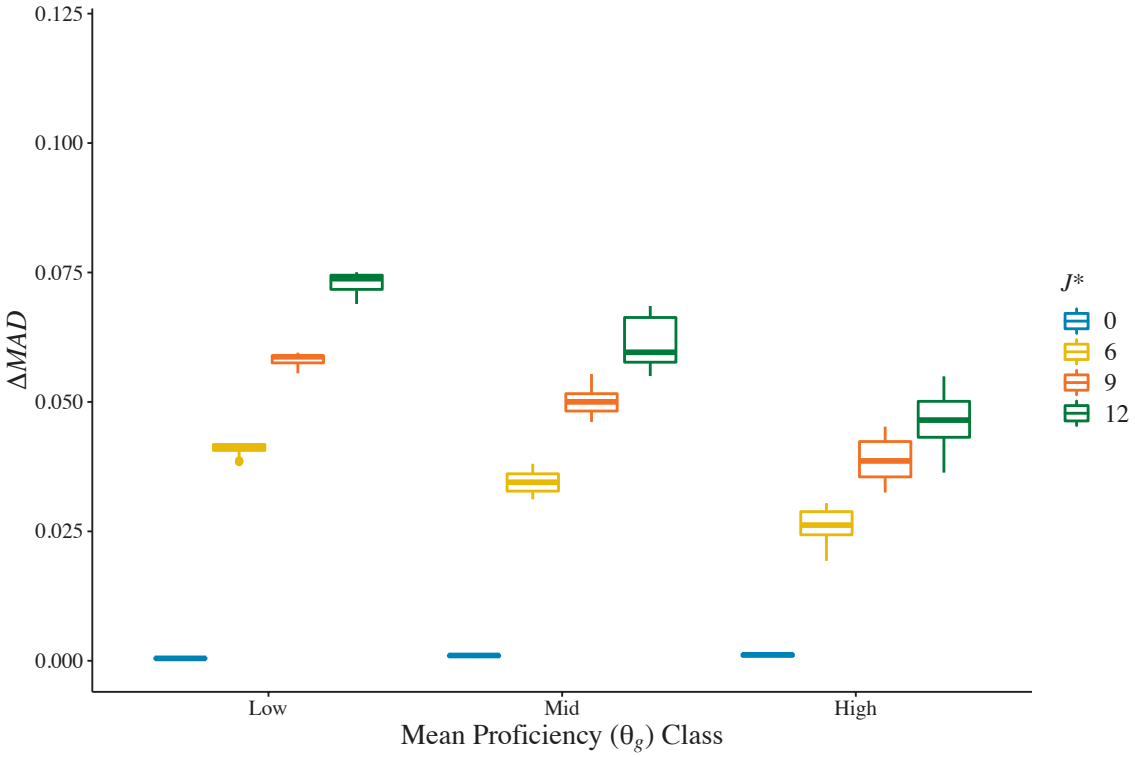


Figure 4.4. Boxplot Comparison of *ΔMAD* by Number of Target Items (*J**) and Mean Proficiency Class for Study 2B-SH

GRF plots of minimum and maximum adjacent conditions were consistent with the boxplot comparison results. Figures B.10, B.11, and B.12 provide adjacent plots (adjacent levels of subgroup homogeneity) for Low, Mid, and High θ_g classes, respectively. In the Low θ_g class, all adjacent plots had clear differences in separation between target and non-target GRFs, and in the expected direction (i.e., minimum plots had greater separation than maximum plots). For Mid and High θ_g classes, differences in adjacent plots for $J^* = 9$ and $J^* = 12$ were unclear, which is consistent with boxplot overlap for the same conditions. Non-adjacent plots for the same conditions (i.e., *minimum* plots for $J^* = 9$ and $J^* = 12$) show clear separation, which is also consistent

with boxplots. Compared to the previous sub-study, the adjacent plots in this sub-study tend to show clearer visual differences in subset separation. Again, this effect appears to be related to the magnitude of the difference in mean ΔMAD between adjacent conditions. The average difference for this sub-study was 0.016 (*Min* = 0.001, *Max* = 0.025). Whereas the average difference for 2B-GUESS was 0.004 (*Min* = -0.015, *Max*=0.043).

Relationship with ΔMAD_R. As previously discussed, differences in rates of *MAD* change between subsets (target and non-target) are captured by ΔMAD_R but ignored by ΔMAD. As a result, although changes in ΔMAD trended in the hypothesized direction for all sub-studies (i.e., increasing trend), decreasing trends were observed in ΔMAD_R across some conditions. Averaging over θ_g class, mean ΔMAD_R fluctuated between increasing and decreasing as number target items moved from 6 to 12. An example of this behavior is in sub-study 2B-GUESS (Table 4.4). The marginal mean ΔMAD_R decreased from 2.216 to 1.962 as level increased from 6 to 9 target items (decrease of 0.254). However, from 9 to 12 items, ΔMAD_R increased to 2.158 (increase of 0.196). Table C.3 in Appendix C contains the means for each condition for sub-study 2B-GUESS, including *MAD* means for target and non-target subsets. In conditions involving 6 to 12 items, percent *MAD* increase between conditions was up to 9% greater for the non-target subset.

For sub-studies involving subgroup homogeneity (Factor A), the same effect was *not* observed. In these sub-studies, mean ΔMAD_R increased as levels of the factor increased. Recall that number of target items was held at a constant (12 target items). As

a result, the non-target empirical GRF can only move closer to the empirical (i.e., decrease in *MAD*) as level of homogeneity increases. Differences in ΔMAD_R across conditions are amplified relative to differences in ΔMAD because *MAD* for the non-target subset must decrease as *MAD* for the target subset increases.

Research Question 2.4: Results Related to External Criteria

Spearman's rank correlations (r_s) are provided in Tables 4.5 and 4.6 for sub-studies involving guessing (GUESS) and spuriously high responding (SH), respectively. Correlations are between ΔMAD and the mean person fit statistics for the subgroup (averaged over 100 subgroup members and 50 replications of the data; e.g., see Equation 16). Correlations with ΔMAD_R are also provided. Correlations in these tables are based on 600 subgroups. Again, person fit statistics considered were outfit mean square (*U*; Equation 12), infit mean square (*W*; Equation 14), and between-fit mean square (*UB*; Equation 15), which accounts for differences in fit between subsets. Table C.5 in Appendix C provides mean person fit statistics averaged for each condition in each sub-study.

Table 4.5. Spearman's Rank Correlations (n = 600) for Guessing Sub-Studies (Study 2)

Sub-Study	Factor	Mean Fit Statistic	Correlation w/	
			ΔMAD	***ΔMAD$_R$***
2A-GUESS	*K**	U_M	0.253	-0.247
		W_M	0.415	-0.104[a]
		UB_M	0.622	0.139
2B-GUESS	*J**	U_M	0.922	0.685
		W_M	0.949	0.755
		UB_M	0.989	0.768

Note. Each correlation is composed of n = 600 simulees. *ΔMAD* is the difference in *MAD* values between subset GRFs; ΔMAD_R is the ratio of *ΔMAD* to *MAD* for the non-target subset; U_M is the mean outfit statistic for the subgroup, W_M is the mean infit statistic for the subgroup, UB_M is mean between-fit statistic for the subgroup; *K** is subgroup homogeneity factor (Factor A); *J** is the number of target items (Factor B).

[a] $p < 0.05$. $p < 0.001$ for all other correlations.

Sub-study 2A-GUESS. When considering the effect of random guessing and variation in subgroup homogeneity, correlations between fit statistics and *ΔMAD* were low to moderate and positive. The largest correlation was with UB_M (r_s = 0.622; df = 598; $p < 0.001$). The lowest correlation was with U_M (r_s = 0.253; df = 598; $p < 0.001$).

Sub-study 2B-GUESS. When considering the effect of random guessing and variation in number of target items, correlations between fit statistics and *ΔMAD* were very high and positive. The largest correlation was with UB_M (r_s = 0.989; df = 598; $p <$ 0.001). The lowest correlation was with U_M (r_s = 0.922; df = 598; $p < 0.001$).

Table 4.6. Spearman's Rank Correlations (n = 600) for SH Sub-Studies (Study 2)

Sub-Study	Factor	Mean Fit Statistic	Correlation w/ ΔMAD	Correlation w/ ΔMAD_R
2A-SH	K*	U_M	0.323	0.267
		W_M	0.327	0.269
		UB_M	0.592	0.541
2B-SH	J*	U_M	0.583	0.592
		W_M	0.519	0.564
		UB_M	0.981	0.590

Note. Each correlation is composed of n = 600 simulees. All correlations are significant at the 0.001 level. ΔMAD is the difference in MAD values between subset GRFs; ΔMAD_R is the ratio of ΔMAD to MAD for the non-target subset; U_M is the mean outfit statistic for the subgroup, W_M is the mean infit statistic for the subgroup, UB_M is mean between-fit statistic for the subgroup; K* is the subgroup homogeneity factor; J* is the number of target items factor.

Sub-study 2A-SH. When considering the effect of spuriously high responding and variation in subgroup homogeneity, correlations between fit statistics and ΔMAD were low to moderate and positive. The largest correlation was with UB_M (r_s = 0.592; df = 598; p < 0.001). The lowest correlation was with U_M (r_s = 0.323; df = 598; p < 0.001).

Sub-study 2B-SH. When considering the effect of simulated guessing and variation in number of target items, correlations between fit statistics and ΔMAD were moderate to very high and positive. The largest correlation was with UB_M (r_s = 0.981; df = 598; p < 0.001). The lowest correlation was with W_M (r_s = 0.519; df = 598; p < 0.001).

Correlations with ΔMAD_R. With a few exceptions, correlations with ΔMAD_R tended to be weaker than those involving ΔMAD. This result is expected given that

ΔMAD_R has been demonstrated to fluctuate in direction (increasing and decreasing) over increasing levels of the main factor due to differential rates of change in *MAD* between non-target and target subsets. Across sub-studies, the most extreme positive correlation was with UB_M ($r_s = 0.768$; $df = 598$; $p < 0.001$) in sub-study 2B-GUESS.

Study 2 Conclusion

Results involving internal criteria provide evidence that GRFs are sensitive to changes in aberrance severity. Trends in *ΔMAD* were as expected, considering either marginal means or boxplot comparisons. As either subgroup homogeneity or number of target items increased, *ΔMAD* increased. These increases were most significant for conditions involving (1) Low θ_g subgroups and spuriously high responding; or (2) High θ_g subgroups and guessing. Differences in distribution of *ΔMAD* between conditions (as observed in boxplots) were generally reflected in adjacent GRF plots. The GRF plots also provide evidence that the GRFs are sensitive to differences in the nature (or type) of aberrance. As expected, empirical GRFs involving substantial guessing generally fell below the theoretical GRF, while those involving substantial spuriously high responding fell above the theoretical GRF.

Results involving external criteria also provide evidence of GRF sensitivity. Most correlations between mean person fit statistics and *ΔMAD* were moderate to high. Over sub-studies, *ΔMAD* correlated highest with UB_M (*Min* = 0.592; *Max* = 0.989). This result is expected because both *ΔMAD* and *UB* are both local measures of fit, whereas *W* and *U* are global measures. Correlations with *ΔMAD* were largest, on average, for aberrance due

to guessing than for spuriously high responding. Also, correlations with *ΔMAD* were largest, on average, for aberrance related to variations in number of target items as opposed to variation in subgroup homogeneity.

Study 3 Results

As discussed previously, Study 3 involved several steps, including preliminary analyses and GRF analyses. Preliminary analyses involved obtaining various fit indices for each examinee and comparing groups (EL versus Non-EL) on fit distribution. There were nine GRF analyses, composed of six misfit conditions (no flags; flagged on *U;* flagged on *W*; flagged on $UB_{diff;}$ flagged on $UB_{ord;}$ flagged on UB_{pty}) and three subset comparisons: item order (ORD), passage type (PTY), and difficulty classification (DIFF). Flagged examinees were grouped into EL or Non-EL and their GRF plots were compared for each of the nine analyses.

Results described here address research questions 3.1 and 3.2. Although the results presented here correspond to the research questions, results are organized by Analysis for a more conceptually straightforward presentation and in keeping with the analytical process described in Chapter 3. However, in the discussion, research questions will be addressed distinctly.

Total number of EL examinees in the data was 2,919 (50,246 Non-EL examinees). Number of examinees composing each GRF plot depended on language classification (EL subgroups were smaller than Non-EL subgroups) and fit index used. For EL plots, the minimum subgroup size across analyses was 32 (approximately 1% of EL examinees in the data). The maximum was 333 (approximately 11%). For Non-EL

plots, the minimum was 153 (0.3% of the total Non-EL). The maximum was 1,922 (approximately 4% of total Non-EL). The smoothing iteration guidelines established in Study 1 were applied.

Results of Preliminary Analyses

Table 4.7 provides subgroup comparisons of summary statistics for proficiency estimate distribution and distribution of each person fit index. Figure D.1 in Appendix D contains the corresponding density plots, comparing subgroups on each index. EL examinees were on average estimated as less proficient than Non-EL students (an average difference of 1.208 logits). For each fit index, compared to Non-EL examinees, EL examinees were on average more misfitting and more varied in distribution of fit. The most noticeable difference was for UB_{diff}. Mean UB_{diff} was 1.960 (SD = 3.096; $Q3$ = 2.238) for EL examinees and 1.270 (SD = 1.929; $Q3$ = 1.158) for Non-EL examinee.

For a given fit index, examinees with values more extreme than the 95th percentile critical value computed from the bootstrap sampling procedure were classified as aberrant. Critical values can be found in Table 4.8. Table 4.8 also provides comparisons of aberrant examines, by subgroup, on each fit index and proficiency estimate. Figure D.2 in Appendix D contains the corresponding density plot comparisons (i.e., comparing aberrant examinees, by subgroup, on distribution of each index).

For all but UB_{ord}, EL examinees were flagged at a higher rate (proportion of total subgroup in the data) than Non-EL examinees. The largest difference in proportion (P) was for U (P_{EL} = 0.114; P_{NEL} = 0.038). Among the between fit indices, the largest difference in proportion was for UB_{diff} (P_{EL} = 0.085; P_{NEL} = 0.030). Flagging rate for the

total sample (including both subgroups) was 0.042 for U and 0.033 for UB_{diff}. Across all fit indices, compared to flagged Non-EL examinees, flagged EL examinees were on average more misfitting and had lower proficiency estimates.

Table 4.7. Summary Statistics for All Examinees by Subgroup (Study 3)

Statistic	**Subgroup**	***Mean***	***SD***	***MIN***	***Q1***	***Q3***	***MAX***
$\hat{\theta}$	Non EL	1.005	1.176	-2.650	0.170	1.830	5.460
	EL	-0.203	0.915	-2.400	-0.920	0.370	5.460
U	Non EL	0.978	0.292	0.006	0.809	1.111	6.633
	EL	1.164	0.326	0.006	0.943	1.307	3.736
W	Non EL	0.987	0.131	0.014	0.898	1.070	1.633
	EL	1.059	0.133	0.014	0.966	1.143	1.549
UB_{diff}	Non EL	1.270	1.929	0.000	0.177	1.581	36.269
	EL	1.960	3.096	0.000	0.178	2.238	25.421
UB_{ord}	Non EL	1.178	1.661	0.000	0.157	1.502	29.352
	EL	1.234	1.763	0.000	0.164	1.721	18.552
UB_{pty}	Non EL	1.178	1.632	0.000	0.127	1.655	22.570
	EL	1.280	1.791	0.000	0.124	1.749	15.118

Note. $\hat{\theta}$ is the proficiency estimate; U is the outfit index, W is the infit index, UB_{diff} is between fit by item difficulty, UB_{ord} is between fit by item order, UB_{pty} is between fit by passage type. ORD refers to GRF comparison by item order (first and last half of the test), PTY refers to GRF comparison by passage type (literary versus informational), and DIFF refers to GRF comparison by difficulty (item difficulty estimates either less than or greater than 0 logits).

Table 4.8. Summary Statistics for Flagged Examinees by Subgroup (Study3)

Fit Index (95th Pctl)	**Subgroup**	**Summary of Fit Index**				**Proficiency ($\hat{\theta}$)**	
		N	***P***	***Mean***	***SD***	***Mean***	***SD***
U	Non EL	1922	0.038	1.839	0.318	-0.177	1.732
(1.549)	EL	333	0.114	1.852	0.272	-1.100	0.518
	Total	2255	0.042	1.841	0.311	-0.314	1.645
W	Non EL	153	0.003	1.415	0.039	-0.498	0.661
(1.371)	EL	32	0.011	1.427	0.045	-0.771	0.321
	Total	185	0.003	1.417	0.040	-0.546	0.623
UB_{diff}	Non EL	1498	0.030	9.166	3.572	-0.387	1.060
(5.954)	EL	247	0.085	10.322	4.077	-0.939	0.593
	Total	1745	0.033	9.330	3.669	-0.465	1.026
UB_{ord}	Non EL	1155	0.023	8.154	2.348	0.508	0.985
(5.876)	EL	62	0.021	8.903	2.871	-0.271	0.684
	Total	1217	0.023	8.192	2.382	0.469	0.987
UB_{pty}	Non EL	1214	0.024	7.910	1.968	0.654	1.112
(5.924)	EL	93	0.032	8.064	2.137	-0.577	0.738
	Total	1307	0.025	7.921	1.980	0.567	1.135

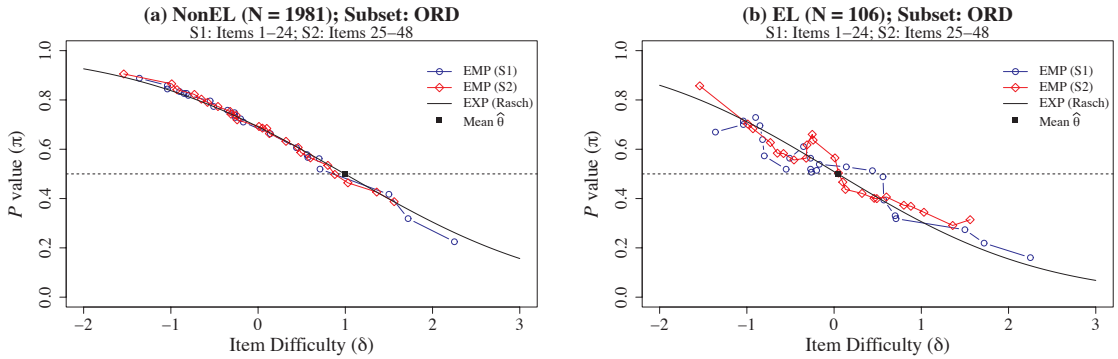
Note. 95th Pctl is the upper bound critical value for flagging aberrant examinees; *N* is the number of flagged examinees; *P* is the proportion of flagged examinees; *U* is the outfit index, *W* is the infit index, UB_{diff} is between fit by item difficulty, UB_{ord} is between fit by item order, UB_{pty} is between fit by passage type. ORD refers to GRF comparison by item order (first and last half of the test), PTY refers to GRF comparison by passage type (literary versus informational), and DIFF refers to GRF comparison by difficulty (item difficulty estimates either less than or greater than 0 logits).

Results of GRF Analyses

Analyses 1 and 2. Analyses 1 and 2 involved GRF comparisons for examinees who were not flagged on any of the fit indices, which provided a baseline comparison for subsequent analyses. Figure 4.5 provides GRF plots for Analyses 1 and 2. A random sample of 10% of examinees in each subgroup was used for these analyses. The same sample was used in each analysis (N_{EL} = 106; N_{NEL} = 1,981). Analysis 1 subset items by order (ORD) and Analysis 2 subset items by passage type (PTY). For both analyses and for both subgroups, both subset GRFs were approximately monotonic decreasing, generally consistent with the theoretical (or expected) GRF, and consistent with the average proficiency estimate for the subgroup (i.e., crossing $\pi(\delta) = 0.50$ near the expectation).

Compared to Non-EL, GRF plots for EL are noticeably less smooth. Comparable GRF plots generated in Study 2 (i.e., for 100 simulees and zero target items), tend to appear smoother and more consistent with the theoretical GRF. Lack of smoothness for the EL GRFs, especially in Analysis 1 (ORD), may therefore also suggest another (unknown) item characteristic influencing aberrance. Also, as previously noted, EL examinees were on average more misfitting than Non-EL on each of the studied fit indices. Although these examinees were not flagged as aberrant, misfit exists to some degree, and therefore some inconsistency with the theoretical GRF may be expected.

ANALYSIS 1: NO MISFIT FLAGS; SUBSET BY ITEM ORDER (ORD)



ANALYSIS 2: NO MISFIT FLAGS; SUBSET BY PASSAGE TYPE (PTY)

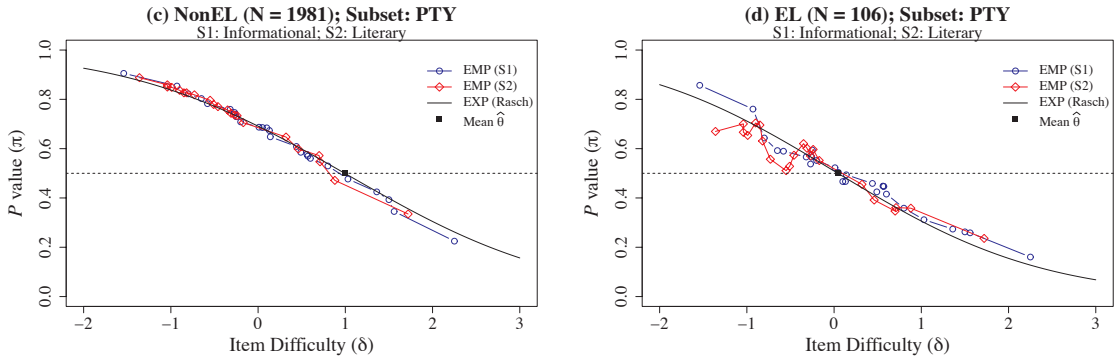
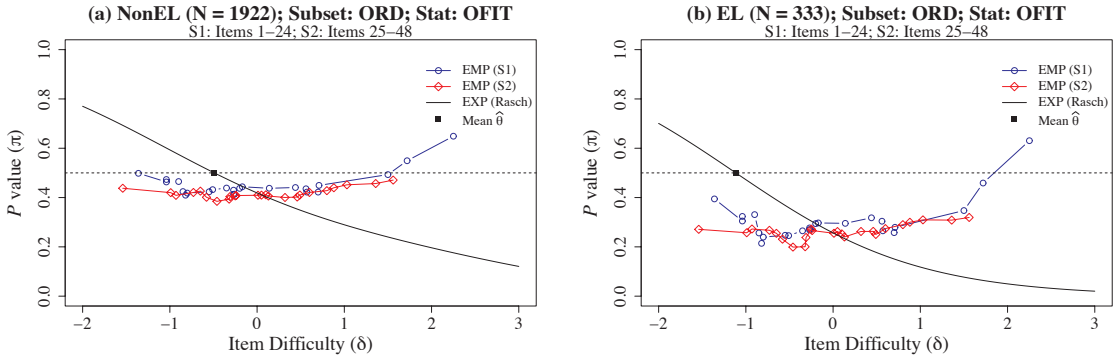


Figure 4.5. GRF Analyses 1 and 2 (Study 3). Empirical (EMP) and Expected (EXP) GRFs for Random 10% of Unflagged Examinees by Item Subset Type and Subgroup

Analyses 3 and 4. Analyses 3 and 4 involved GRF comparisons for examinees who were flagged on *U*. Figure 4.6 provides GRF plots for Analyses 3 and 4. Analysis 3 subset items by order (ORD) and Analysis 4 subset items by passage type (PTY). GRF patterns were very similar between analyses for a given subgroup. Also, subset GRFs for a given analysis and given subgroup were very similar. These similarities suggest that neither item order nor passage type had much, if any, influence on aberrance for these examinees.

ANALYSIS 3: FLAGGED ON OUTFIT (*U*); SUBSET BY ITEM ORDER (ORD)



ANALYSIS 4: FLAGGED ON OUTFIT (*U*); SUBSET BY PASSAGE TYPE (PTY)

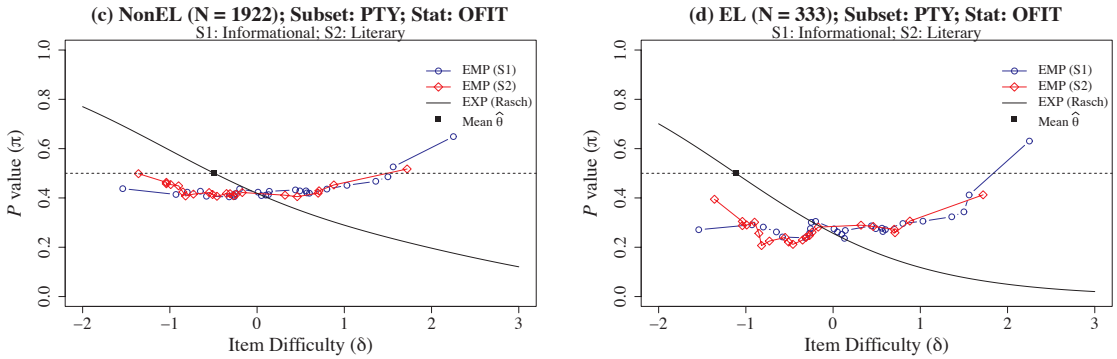
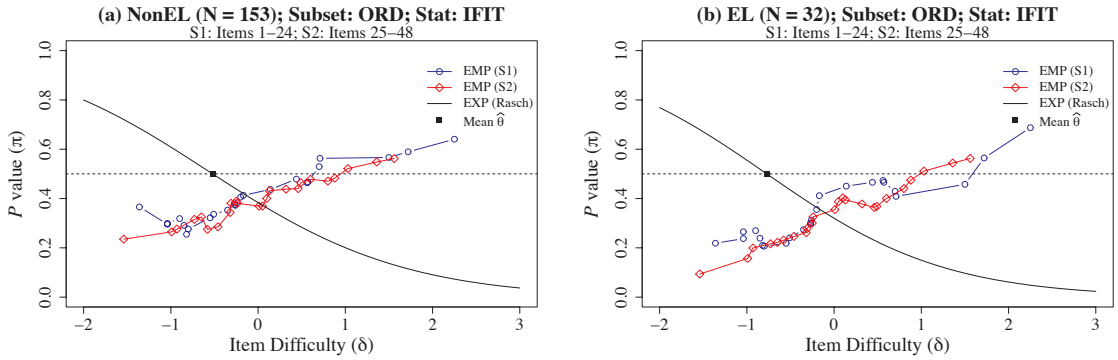


Figure 4.6. GRF Analyses 3 and 4 (Study 3). Empirical (EMP) and Expected (EXP) GRFs for Examinees Flagged on OutFit (*U*) by Item Subset Type and Subgroup

For both subgroups, GRF patterns are slightly U-shaped. Near the low end of the difficulty spectrum, GRFs are slightly decreasing. Near the high end of the spectrum, GRFs increase. Such patterns may indicate scores for these examinees are spuriously high, which may be due to, for example, answer copying on the most difficulty items. However, much of the increase appears to be due to the influence of only one or two

items (the most difficult few items). Patterns are otherwise relatively flat or horizontal, which suggests guessing behavior.

ANALYSIS 5: FLAGGED ON INFIT (*W*); SUBSET BY ITEM ORDER (ORD)



ANALYSIS 6: FLAGGED ON INTFIT (*W*); SUBSET BY PASSAGE TYPE (PTY)

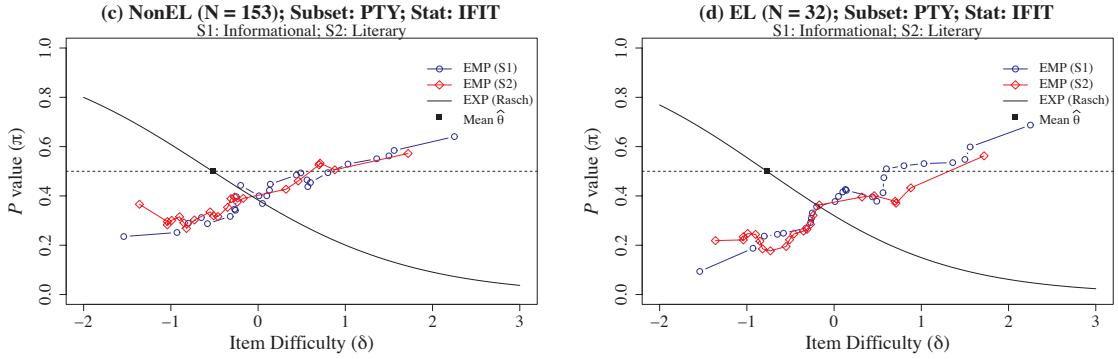
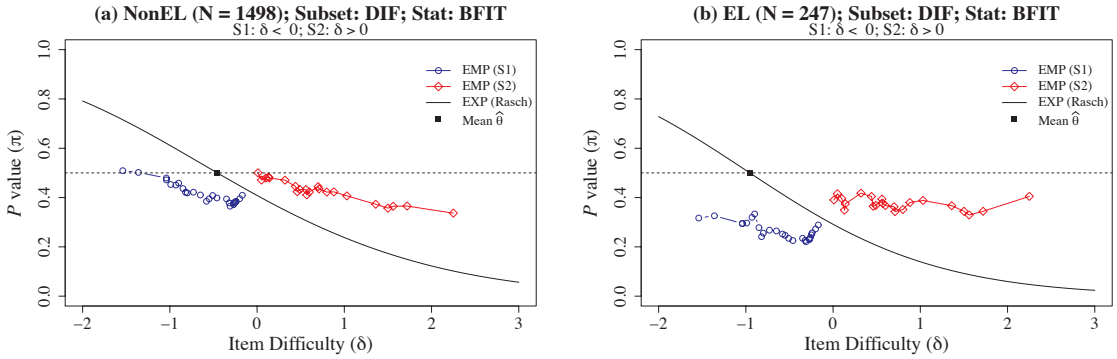


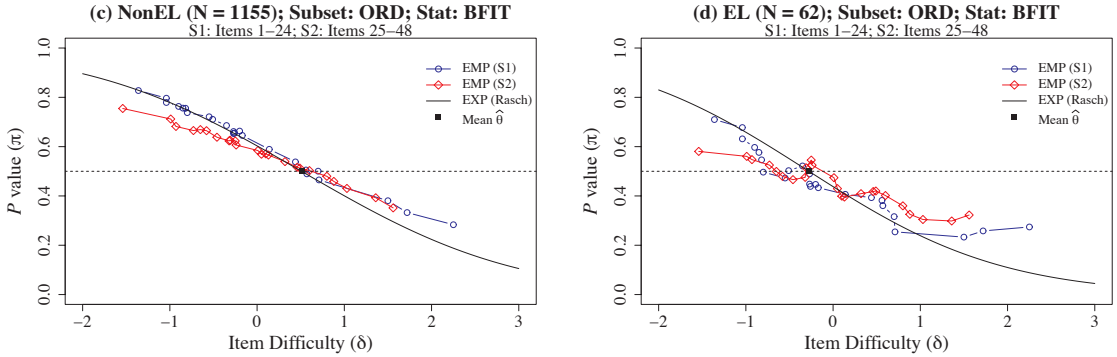
Figure 4.7. GRF Analyses 5 and 6 (Study 3). Empirical (EMP) and Expected (EXP) GRFs for Examinees Flagged on InFit (*W*) by Item Subset Type and Subgroup

Analyses 5 and 6. Analyses 5 and 6 involve GRF comparisons for examinees who were flagged on *W*. Figure 4.7 provides GRF plots for these analyses. Analysis 5 subset items by order (ORD), and Analysis 6 subset items by passage type (PTY). Patterns were very similar between subgroups. Compared to patterns in Analyses 3 and 4,

ANALYSIS 7: FLAGGED ON BETWEEN FIT DIFFICULTY (UB_{diff})



ANALYSIS 8: FLAGGED ON BETWEEN FIT ITEM ORDER (UB_{ord})



ANALYSIS 9: FLAGGED ON BETWEEN-FIT PASSAGE TYPE (UB_{pty})

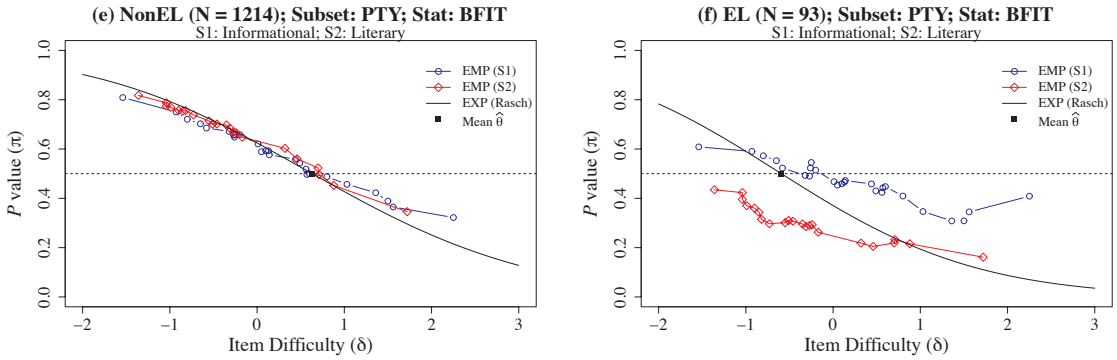


Figure 4.8. GRF Analyses 7-9 (Study 3). Empirical (EMP) and Expected (EXP) GRFs for Examinees Flagged on Between Fit (*UB*) by Item Subset Type and Subgroup

Analyses 5 and 6 were less U-shaped and increased at a greater rate. All patterns were generally monotonic *increasing*, which is the opposite of expectation. *P*-value tended to be near guessing level for easy items but increased rapidly. The rate of increase was greater than the corresponding plots in Analyses 3 and 4, which were much flatter at the middle of the difficulty spectrum. Recall that horizontal or near-horizontal patterns are associated with guessing-type behavior. In contrast, patterns in Analyses 5 and 6 had no noticeable horizontal component. This may be expected given that *W* deemphasizes residuals where there are large differences between examinee proficiency and item difficulty. Examinees may be more likely to guess on items that are relatively (relative to their proficiency level) more difficult. *U* would be more likely than *W* to capture this behavior.

Analyses 7-9. Analyses 7 through 9 involved GRF comparisons for examinees who were flagged on between fit indices. Figure 4.8 provides GRF plots for these analyses.

Analysis 7 compared subgroups on between fit by item difficulty, UB_{diff}. For both subgroups, the *easy* item subset fell below the expected GRF and the *difficult* subset fell above. The overall pattern, considering both subsets, was relatively horizontal, suggesting guessing behavior. Although similar in shape, the GRFs for the EL subgroup were more horizontally oriented and lower in *p*-value, on average, compared to Non-EL. This effect is likely due to differences in average proficiency between the two groups. The EL subgroup had a mean proficiency estimate of -0.939 compared to -0.387 for the Non-EL subgroup (see Table 4.8).

Analysis 8 compared subgroups on between fit by item order, UB_{ord}. Although flagged as aberrant, the patterns suggest that the level of validity threat to score interpretations for these examinees is relatively low (e.g., compared to examinees in Analyses 3-6). For both subgroups, both subset GRFs are generally monotonic decreasing and cross relatively close to their respective average proficiency estimates. However, for both subgroups, for easier items, the GRF associated with the last half of the test was less consistent with expectation than the GRF associated with the first half of the test. As previously discussed, examinees who run out of time toward the latter half of a test may resort to rapid guessing for the remaining items. For the EL subgroup, differences between subset GRFs can also be seen for difficult items. This was not the case for the Non-EL subgroup. The interaction may again be due to differences in average proficiency between the two groups (a difference of 0.779 logits). As EL examinees were on average estimated as less proficient, compared to Non-EL examinees, guessing on more difficult items may have provided more of an advantage in terms of probability of correct response.

Compared to all other analyses in Study 3, results from Analysis 9 most clearly demonstrated a subgroup by subset interaction with respect to aberrance. For EL examinees, *informational* and *literary* GRFs had clear separation, whereas the same GRFs for the Non-ELL group had little to no separation. The difference was striking considering the two subgroups were flagged on the same index, using the same flagging criteria, and are on average aberrant to a similar degree. The mean fit statistic was 7.910 (*SD* = 1.968) for flagged Non-EL examinees and 8.064 (*SD* = 2.137) for flagged EL

examinees (Table 4.8). Difference in average ability estimate between subgroups was the largest of all the analyses (1.231 logits).

For the Non-EL subgroup, both subset GRFs were relatively consistent with the expected GRF and crossed close to the average proficiency estimate. This result suggests that although members of the Non-EL subgroup were flagged as aberrant, severity of aberrance at the subgroup level was relatively mild. For the EL subgroup, both subset GRFs are generally monotonic decreasing. However, the *literary* GRF was much lower than expectation (examinees performed worse than expected), and the *informational* GRF was much higher than expectation (examinees performed better than expected). The *informational* GRF was also much closer to the average proficiency estimate than the *literary* GRF, which suggests that performance on *informational* subset is a relatively better indicator of subgroup proficiency.

Study 3 Conclusion

Comparing across fit statistics, GRF pattern variation (e.g., U-shaped, horizontal) suggests the GRF approach is useful for discriminating between different aberrance-related response behaviors, including guessing behavior and potential preknowledge or copying behavior. GRF results for EL and Non-EL were relatively similar across analyses with the exception of Analysis 9. In Analysis 9, aberrance was more severe for the EL subgroup, as indicated by greater inconsistency between subset GRFs and greater inconsistency with the expected GRF. Analysis 9 for the EL group was the only analysis for which subset GRFs had clear separation, which represents a clear difference in response behavior for the two subsets.

CHAPTER V

DISCUSSION

Comprehensive approaches to person fit combine global, local, and graphical analyses to contextualize aberrant test responses. Such approaches may provide insight into the severity and nature of aberrance, which, in turn, may speak to potential fairness and validity issues. Three studies were conducted as an initial exploration of the use of group response functions (GRFs) to contextualize aberrant responses in a K-12 educational testing context. The studies addressed the *accuracy*, *sensitivity*, and *practicality* (respectively) of the proposed GRF approach. The approach compares GRFs both between item subsets and between examinee subgroups. Differential GRF patterns suggest aberrance related to item characteristics, subgroup characteristics, or the interaction between item and subgroup characteristics.

Simulation Studies

Studies 1 and 2 used simulated data under relatively ideal conditions to provide a "proof of concept" of the approach.

Study 1

Study 1 investigated the accuracy of empirical GRFs as estimates of the expected, theoretical GRF. In particular, research questions 1a and 1b address the impact of subset and subgroup size on number of smoothing iterations. In general, subset and subgroup size were found to have very little influence. Over all conditions, the average error in

estimation was relatively low *without* smoothing, particularly for subgroup sizes greater than 100. This result suggests that in most conditions, the classical *p*-value (the proportion of subgroup members responding correctly to an item) is sufficient to construct the GRF. However, at least one smoothing iteration improved accuracy somewhat for most conditions. Conditions of 24 items and 100 or fewer simulees benefited from more than one iteration.

Compared to 24-item conditions, bias increased at a greater rate over iterations for the 12-item conditions. As a result, 12-item conditions benefited less from smoothing. This difference has implications for potential future studies involving item subset sizes more comparable to those found in passage-based tests (e.g., 6 to 9 items). In such conditions, rapid increase in bias may offset any effort to reduce overall error via smoothing. This disadvantage may be inconsequential for larger subgroup sizes (e.g., greater than 100), as the unsmoothed GRF under these conditions may be a sufficient estimate. For the present study, real data analyses involved item subsets that were approximately 24 items and subgroup sizes between 32 and 1,922. Results from Study 1 provided guidelines for selecting an appropriate number of smoothing iterations.

Study 2

Study 2 provided evidence of GRF sensitivity to variation in aberrance severity and nature. There were four research questions. Research question 2.1 addressed the extent to which homogeneity of aberrant responding in the subgroup affects GRF separation between item subsets. As homogeneity increased, separation between subset GRFs increased. Research question 2.2 addressed the extent to which number of target

items affects GRF separation between item subsets. As number of target items increased, separation between subset GRFs increased. Research question 2.3 addressed the extent to which type of aberrance affects GRF separation between item subsets. For the guessing condition, separation was greatest for high proficiency subgroups. For the spuriously high condition, separation was greatest for low proficiency subgroups. Research question 2.4 addressed the extent to which GRF separation corresponds to other available measures of person fit. In general, separation had moderately low to high correspondence to available measures. Correspondence was highest with measures that consider differences in fit between different item subsets. These conclusions are supported by the technical details of Study 2 results, which are summarized below.

Trends in internal criteria, particularly *ΔMAD*, were in the hypothesized direction: increasing as either subgroup homogeneity or number of target items increased. These differences were confirmed in GRF plot comparisons of adjacent conditions. The exception was for adjacent conditions where the difference in *ΔMAD* was relatively small (e.g., < 0.01). Non-adjacent conditions generally showed clear differences (in boxplots, as well as in GRF plot comparisons). Differences were greatest for (1) low proficiency subgroups when spuriously high responding was simulated and (2) high proficiency subgroups when random guessing was simulated. These results are as expected. In terms of increasing probability of a correct response, low proficiency examinees would benefit more from spurious increases in proficiency than high proficiency examinees. Similarly, high proficiency subgroups would benefit less from random guessing than low proficiency examinees.

Spuriously high (SH) responses and guessing were clearly differentiable in GRF plots. Guessing GRFs generally fell below the expectation; SH GRFs generally fell above expectation. However, Study 2 represented the ideal condition in which only one or the other aberrance type was present. In addition, the guessing condition was random guessing. However, examinees may use other guessing-type strategies. For example, some examinees in Walker et al (2016) reported using "cued" guessing. PRFs for these examinees were more consistent with expectation compared to those who reported using random guessing only. In less than ideal conditions, when aberrance-related behaviors and response strategies are mixed between and within examinees, the GRF approach may be less sensitive and less able to make distinctions between different types of behaviors.

In terms of external criteria, correlations of average fit indices with *ΔMAD* were between 0.25 and 0.99. On average, correlations were higher for guessing than for SH and higher when manipulating number of target items, as opposed to subgroup homogeneity. With the exception of sub-study 2B-SH, correlations were highest for UB_M and lowest for U_M. This result is not surprising given that, unlike *U* and *W*, both UB_M and *ΔMAD* are designed to account for differences in aberrance between two disjoint item subsets. Taken together, these results provide evidence that the GRF approach is in most conditions at least moderately consistent with established fit indices.

Real Data Study

Study 3 provided a real data demonstration of the GRF approach. Nine analyses were conducted on each of two subgroups (EL and Non-EL examinees). To contextualize

aberrant responding, analyses were compared between subgroups and between item subsets for a given subgroup.

Research question 3.1 relates to comparisons between item subsets for a given subgroup. Results across all nine analyses show that the GRF approach is clearly able to produce subset GRFs that differ in shape and orientation and that are consistent with traditional measures of person fit. Research question 3.2 relates to comparisons of GRF patterns between subgroups and the ability of the GRF approach to detect seemingly important interactions between subgroup and subset with respect to aberrance. For some analyses, GRF patterns were very similar between subgroups. For other analyses, GRF patterns differed. This suggests that the GRF approach is useful for detecting subgroup-based differences. In particular, Analysis 9 showed a substantial difference between subgroup patterns grouped by passage type.

For analyses involving examinees flagged on global indices (*U* and *W*), few differences were observed between subgroup GRF plots. Also, *within* GRF plots for each subgroup, GRFs for each item subset were similar, having approximately the same shape and orientation. Some slight differences appear to be related to consistently lower average proficiency estimates for EL subgroups. *P*-values tended to be lower for EL subgroups, but again, the basic patterns for the subgroups are the same and the effect seems to be consistent over item subsets. The overall result suggests that subgroups flagged on global indices are similar in aberrance severity and type. This result provides no information on item or subgroup characteristics that may underlie aberrant responding.

For analyses involving examinees flagged on between fit indices, GRF patterns were similar between subgroups when items were subset by either order or difficulty (Analyses 7 and 8, respectively). Although patterns were similar, GRF plots for EL subgroups appeared to be more inconsistent with the expected GRF, suggesting greater severity of aberrance, and therefore greater validity threat, for the EL subgroup.

In Analysis 9, although both subgroups were flagged on UB_{pty} and had similar UB_{pty} distributions, very little difference between subset GRFs was observed for the Non-EL subgroup. Subset GRFs were also relatively consistent with the expected GRF for the Non-EL group. In contrast, for the EL subgroup, subset GRFs showed clear separation and were relatively inconsistent with expectation. EL examinees performed lower than expected on items based on *literary* passages and higher than expected on items based on *informational* passages. The EL subgroup was estimated at a lower average proficiency than the Non-EL subgroup. However, difference in average proficiency estimate does not explain the difference in patterns in terms of the degree of separation between subset GRFs. Although beyond the scope of the present study, a full discussion of this result would require a formal analysis of the differences between literary and informational passages, as well as differences in EL cognitive processes and learning progressions as it pertains to the two content types. Nevertheless, the result clearly shows the promise of the GRF method. Groupings of both examinees and items can be defined *a priori* and substantial visual interactions can indicate where more formal examination is warranted.

As discussed previously, Petridou and Williams (2010) distinguished between "construct-relevant" and "construct-irrelevant" explanations of aberrance. In Study 3, for

example, the construct of interest was English language and reading. Some EL examinees were observed to respond differentially on items based on literary versus informational passages (i.e., Analysis 9). For these examinees, the empirical GRF for the literary subset fell below the expected GRF, which is consistent with guessing-type behavior (for example, see Figures B.1 through B.3 in Appendix B). The GRF for the informational subset fell above the line, which is consistent with cheating-type behaviors (for example, see Figures B.4 through B.6 in Appendix B). Understanding these patterns in terms of "construct-relevant" or "irrelevant" may not be straightforward, however. For example, the effect seems particular to EL students, as the same difference was not observed in Non-ELs. Language learning status is ostensibly related to the construct of interest. Yet, Abedi (2005) notes that differences in non-linguistic (e.g., cultural, socioeconomic) factors may be construct-*irrelevant* sources of differential performance for ELs.

Also, guessing and "cheating" are typically regarded as construct-irrelevant sources. Validity of score interpretations and uses depends, in part, on the degree to which observed responses reflect expected cognitive processes and methods of interacting with test items and tasks (AERA et al., 2014). Guessing and cheating behaviors are examples of unexpected methods of interacting with test tasks. However, neither behaviors are homogenous. For example, examinees with low motivation may guess randomly, whereas other examinees may use "informed" guessing (e.g., Walker et al., 2016), which suggests application of partial knowledge to eliminate distractor options, thus increasing probability of answering correctly. In addition, spuriously high response patterns may be due to cheating (e.g., answer copying) but may also be related

to non-cheating factors. For example, for EL students in Analysis 9, informational passages may have provided more contextual support than literary passages, facilitating examinee responses in unanticipated ways.

In terms of exploring subgroup- and item-based correlates of aberrance, the GRF approach as demonstrated in Study 3 shows some promise. First, compared to PRF approaches, the GRF approach makes exploration of aberrance more practical for large sample sizes with potentially many individual cases of aberrance. Second, the GRF approach brings several levels of information together into one graphic. Differences in GRF patterns and inconsistencies from the expected GRF can be investigated at the *item level* (particularly in terms of item difficulty), *subset level*, and *subgroup level*. Analysis 9 provides the clearest example of a subgroup-by-subset interaction with respect to aberrance, and therefore the clearest example of the value of the method. For examinees flagged on UB_{pty}, response behavior for literary versus informational passages appears to be qualitatively different, but only for examinees classified as English learners. This interaction signals potential fairness and validity issues that may warrant further investigation.

For the EL examinees in Analysis 9, score interpretation should be undertaken with caution, as response patterns for both passage types are unexpected given the theoretical GRF. Meijer and Tendeiro (2014) suggest that regardless of the "underlying mechanism that causes" aberrance, "a test administrator may decide that the total scores of test takers who are classified as aberrant cannot be trusted and that additional information should be obtained" (p. 15). In addition, person-fit analyses can be framed as

a quality control process, which should involve an exploration of the potential causes of aberrance observed in the sample. In the quality control conception of person-fit analyses, an investigation of potential causes may help identify problems with test characteristics "that may threaten test fairness and measurement accuracy" (Cui & Mousavi, 2015, p. 46). Again, terms like "aberrance" and "person misfit" should not be construed to indicate deficits in examinees. As previously discussed, there are many potential sources of aberrance and not all are negative. In addition, some authors have framed examinee responding in terms of interactions between task and person characteristics, which serves to destigmatize response aberrance for examinees. For example, see Mislevy's work on *sociocognitive* assessment frameworks (e.g., Mislevy, 2018) or Shavelson's work on *generalizability theory* (e.g., Shavelson & Webb, 1991).

Results like that of Analyses 9 provide substantive information that may draw attention to content (e.g., literary content) on which particular subgroups (e.g., EL examinees) may need more support. The Analysis 9 result also provides support that the GRF approach represents a meaningful contribution toward the call for more comprehensive person fit methodologies.

Future Considerations

Analyses in the present study included only aberrant examinees. The justification for this was to help explain aberrant responses in terms of relationships with subgroup- or item-subset characteristics, or their interaction. In future studies, a screening approach may be worth exploring. Such an approach would include all subgroup members in the GRF analysis. Such an approach may also fit more readily within the DIF/DPF paradigm.

Variation in expected GRFs between subsets, but within subgroup, suggests that proficiency estimates (for at least some subgroup members) are not invariant over item subsets (i.e., DPF). This approach could also be conceptualized as a pre-screen before conducting formal DPF studies or a method of graphically contextualizing DPF at the subgroup level (e.g., Engelhard et al., 2014).

As previously noted, simulation studies were conducted under relatively ideal conditions. Future studies may explore the effects of scale unreliability on GRF analyses, as well as the impact of including mixed response strategies. For example, in the present study, only random guessing was simulated. Other response behaviors could be simulated to further investigate GRF sensitivity and the ability of the GRF approach to differentiate between behaviors. For example, Karabatsos (2003) simulated five behaviors: cheaters, creatives, lucky guessers, careless, and random guessers. Behavior simulation involved selecting simulees from various ability distributions and manipulating response strings in various ways (e.g., carelessness was simulated as 0.5 probability of incorrect for the 41% easiest items). Note also that that outcomes and interpretations in Study 2 and Study 3 were influenced by results of Study 1. In an expanded simulation, other computational aspects of deriving empirical GRFs could be explored. For example, an expanded simulation may involve a formal exploration of alternative smoothing methods.

Study 3 explored the practicality of the proposed GRF approach by providing a real data demonstration. Study 2 explored the sensitivity of the approach using quantitative measures. Future studies could expand on practicality and sensitivity evidence by collecting judgments from measurement professionals. For example,

simulated GRF plots of various aberrance levels (including a non-aberrance condition) can be shown to a sample of psychometricians. An outcome of interest would be the consistency at which psychometricians can differentiate between aberrant and non-aberrant GRFs. Such studies are critical because the GRF is a graphical tool, which depends on the perceptual judgments of practitioners.

Conclusion

Comprehensive approaches to person-fit analysis require global, local, and graphical analyses. Under relatively ideal simulation conditions, GRFs appear to provide an accurate and sensitive representation of subgroup-level aberrance over different item subsets. Like PRF approaches (e.g., Emons et al., 2004; Walker et al., 2018), the GRF approach explored in this study allows visualization of aberrant response patterns over the difficulty spectrum and in comparison with expectation, given a model. By comparing GRFs (between subgroups or between subsets), judgements can be made regarding relative aberrance severity and nature. Unlike PRF approaches, the GRF approach provides information about subgroup-level aberrance and is practical when investigating large samples of aberrant examinees. A real-data analysis uncovered an interaction showing qualitatively different subset GRF patterns, but only for EL examinees. The interaction calls into question validity of score interpretations for these examinees, and also identifies examinees that may need additional educational support for particular types of content (e.g., literary versus informational content).

REFERENCES

Abedi, J. (2005). Issues and consequences for English language learners. *Yearbook of the National Society for the Study of Education*, *104*, 175–198.

AERA, APA, & NCME (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95–106.

Baker, B. A., Caison, A. L., & Meade, A. W. (2007). Assessing gender-related differential item functioning and predictive validity with the institutional integration scale. *Educational and Psychological Measurement*, *67*, 545–559.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31–44.

Cole, N. (1993). History and development of DIF. In P. W. Holland, H. Wainer, & Educational Testing Service (Eds.), *Differential item functioning*. Lawrence Erlbaum Associates.

Cui, Y., & Mousavi, A. (2015). Explore the usefulness of person-fit analysis on large-scale assessment. *International Journal of Testing*, *15*, 23–49

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland, H. Wainer, & Educational Testing Service (Eds.), *Differential item functioning*. Lawrence Erlbaum Associates.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, *39*, 1–35.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*, 101–119.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*, 585–602.

Engelhard, G., Kobrin, J. L., & Wind, S. A. (2014). Exploring differential subgroup functioning on sat writing items: What happens when English is not a test taker's best language? *International Journal of Testing*, *14*, 339–359.

Engelhard, G. (2015). Hanning (smoothing) of person response functions. *Rasch Measurement Transactions*, *26*, 1392–1393.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133–146.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, H. I. Braun, & Educational Testing Service (Eds.), *Test validity* (pp. 129–145). L. Erlbaum Associates.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298.

Koo, J., Becker, B. J., & Kim, Y. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*, 89–109.

Lamprianou, I., & Boyle, B. (2004). Accuracy of measurement in the context of mathematics national curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement*, *41*, 239–259.

Linacre, J. M. (2016a). *Winsteps® Rasch measurement computer program*.

Linacre, J. M. (2016b). *Winsteps® Rasch measurement computer program user's guide*. Winsteps.com

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, *10*, 217–229.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, *1*, 477–482.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, *90*, 227–238.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement Applied Psychological Measurement*, *25*, 107–135.

Meijer, R. R., & Tendeiro, J. N. (2014). *The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us* (No. 14–03). Law School Admission Council.

Meijer, R. R., & Van Krimpen-Stoop, E. M. L. A. (2001). Person fit across subgroups: An achievement testing example. In A. Boomsma, M. A. J. van van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 377–390). Springer-Verlag.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits a differential item functioning analysis. *Educational and Psychological Measurement*, *69*, 613–635.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417–430.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the *lz* person-fit statistic. *Applied Psychological Measurement*, *22*, 53–69.

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 151–163). Routledge, Taylor & Francis Group.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *JEDM Journal of Educational Measurement*, *44*, 227–247.

Petridou, A., & Williams, J. (2010). Accounting for unexpected test responses through examinees' and their teachers' explanations. *Assessment in Education: Principles, Policy & Practice*, *17*, 357–382.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielson and Lydiche (for Danmarks Paedagogiske Institut).

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3–38.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191–207.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, *45*, 433–444.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*(2), 359–372.

Smith, R. W., & Davis-Becker, S. (2011, April). *Detecting suspect examinees: An application of differential person functioning analysis*. Annual conference of the National Council on Measurement in Education, New Orleans, LA.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* (pp. 83–108). Academic Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co.

Vale, C. D., & Weiss, D. J. (1975). *A study of computer-administered stradaptive ability testing* (Research Report No. 75–4). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267–298.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Duxbury Press.

Walker, A, Engelhard, G., Hedgpeth, M.-W., & Royal, K. (2016). Exploring aberrant responses using person fit and person response functions. *Journal of Applied Measurement*, *17*, 194–208.

Walker, Adrienne, Jennings, J. K., & Engelhard, G. J. (2018). Using person response functions to investigate areas of person misfit related to item characteristics. *Educational Assessment*, *23*, 47–68.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report No. 73–3). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wilson, M. (2008). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates; Taylor & Francis.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Mesa Press.

APPENDIX A

RMSE PLOTS (STUDY 1)

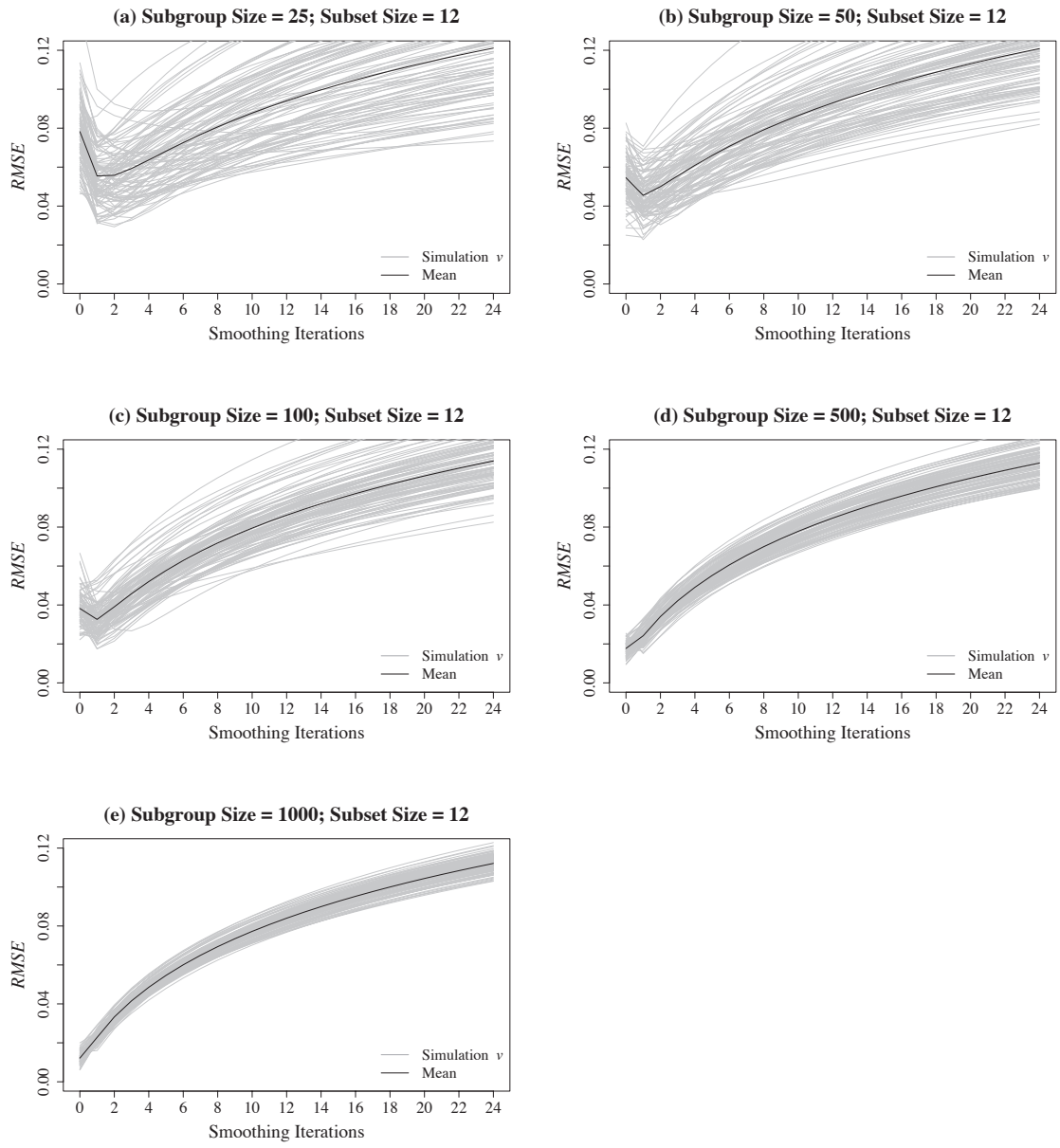


Figure A.1. *RMSE* for 12-item Subset by Subgroup Size and Smoothing Iterations for Each Simulation Replication (v) and Mean *RMSE* over All Replications ($RMSE_M$)

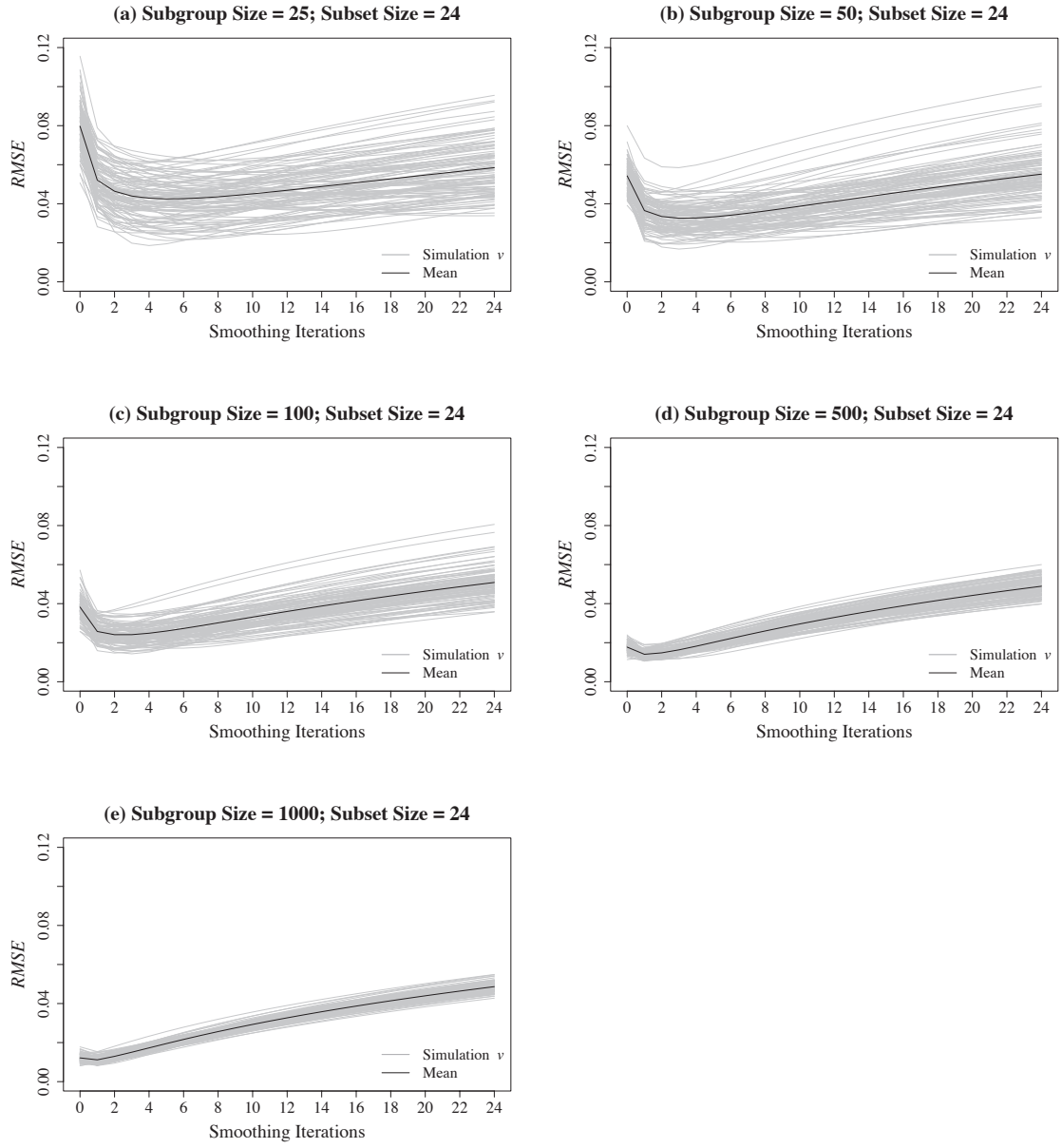


Figure A.2. *RMSE* for 24-item Subset by Subgroup Size and Smoothing Iterations for Each Simulation Replication (v) and Mean *RMSE* over All Replications ($RMSE_M$)

APPENDIX B

MINIMUM AND MAXIMUM GRF COMPARISONS (STUDY 2)

Table B.1. Comparison of Minimum and Maximum *ΔMAD* (Study 2A-GUESS)

Mean Theta Class	***K****	**Status**	**Mean Theta**	MAD_{NT}	MAD_{T}	ΔMAD	ΔMAD_{R}
Low	0	Max	-0.35	0.046	0.060	0.014	0.306
	25	Min	-0.98	0.028	0.035	0.007	0.255
	25	Max	-0.35	0.039	0.066	0.027	0.695
	50	Min	-1.00	0.022	0.034	0.012	0.516
	50	Max	-0.33	0.033	0.070	0.037	1.149
	75	Min	-1.00	0.016	0.037	0.020	1.235
Mid	0	Max	0.32	0.075	0.091	0.016	0.214
	25	Min	-0.31	0.040	0.068	0.027	0.681
	25	Max	0.32	0.062	0.102	0.039	0.629
	50	Min	-0.32	0.033	0.071	0.038	1.144
	50	Max	0.29	0.053	0.109	0.057	1.081
	75	Min	-0.32	0.024	0.078	0.055	2.331
High	0	Max	0.36	0.076	0.092	0.016	0.212
	25	Min	0.34	0.063	0.103	0.040	0.624
	25	Max	0.99	0.089	0.137	0.048	0.541
	50	Min	0.34	0.054	0.112	0.058	1.080
	50	Max	1.00	0.075	0.153	0.078	1.043
	75	Min	0.33	0.038	0.124	0.086	2.245

Note. Max and Min status are, respectively, the maximum and minimum *ΔMAD* for the condition; *K** is group homogeneity (Factor A); GUESS is simulated guessing; *ΔMAD* is the difference in *MAD* values between the target (T) and non-target (NT) subsets; ΔMAD_R is the ratio of *ΔMAD* to *MAD* for the non-target subset.

Table B.2. Comparison of Minimum and Maximum *ΔMAD* (Study 2A-SH)

Mean Theta Class	***K****	**Status**	**Mean Theta**	MAD_{NT}	MAD_{T}	ΔMAD	ΔMAD_{R}
Low	0	Max	-0.98	0.074	0.080	-0.005	-0.068
	25	Min	-0.37	0.083	0.064	0.019	0.302
	25	Max	-0.88	0.088	0.066	0.022	0.335
	50	Min	-0.33	0.098	0.051	0.047	0.928
	50	Max	-0.84	0.103	0.052	0.051	0.970
	75	Min	-0.35	0.108	0.039	0.069	1.746
Mid	0	Max	-0.32	0.069	0.078	-0.009	-0.115
	25	Min	0.31	0.069	0.057	0.012	0.208
	25	Max	-0.25	0.081	0.063	0.019	0.295
	50	Min	0.33	0.081	0.044	0.036	0.820
	50	Max	-0.33	0.097	0.050	0.047	0.934
	75	Min	0.33	0.090	0.035	0.055	1.578
High	0	Max	0.33	0.057	0.069	-0.012	-0.177
	25	Min	0.99	0.050	0.045	0.004	0.098
	25	Max	0.34	0.069	0.057	0.012	0.204
	50	Min	0.95	0.059	0.035	0.023	0.659
	50	Max	0.36	0.079	0.044	0.036	0.810
	75	Min	0.98	0.064	0.028	0.036	1.315

Note. Max and Min status are, respectively, the maximum and minimum *ΔMAD* for the condition; *K** is group homogeneity (Factor A); SH is spuriously high responding; *ΔMAD* is the difference in *MAD* values between the target (T) and non-target (NT) subsets; ΔMAD_R is the ratio of *ΔMAD* to *MAD* for the non-target subset.

Table B.3. Comparison of Minimum and Maximum *ΔMAD* (Study 2B-GUESS)

Mean Theta Class	***J****	**Status**	**Mean Theta**	MAD_{NT}	MAD_T	ΔMAD	ΔMAD_R
Low	0	Max	-0.99	0.007	0.006	0.001	0.129
	6	Min	-0.99	0.018	0.009	0.009	1.040
	6	Max	-0.34	0.037	0.012	0.025	2.098
	9	Min	-1.00	0.026	0.012	0.014	1.173
	9	Max	-0.36	0.054	0.018	0.035	1.918
	12	Min	-1.00	0.037	0.016	0.020	1.235
Mid	0	Max	0.00	0.008	0.007	0.001	0.189
	6	Min	-0.32	0.037	0.012	0.026	2.143
	6	Max	0.33	0.062	0.018	0.044	2.500
	9	Min	-0.32	0.055	0.019	0.036	1.943
	9	Max	0.32	0.090	0.028	0.062	2.252
	12	Min	-0.32	0.078	0.024	0.055	2.331
High	0	Max	0.55	0.008	0.007	0.002	0.220
	6	Min	0.33	0.062	0.018	0.045	2.508
	6	Max	0.99	0.089	0.026	0.063	2.433
	9	Min	0.34	0.091	0.028	0.063	2.240
	9	Max	1.00	0.127	0.040	0.087	2.162
	12	Min	0.33	0.124	0.038	0.086	2.245

Note. Max and Min status are, respectively, the maximum and minimum *ΔMAD* for the condition; *J** number of target items (Factor B); GUESS is simulated guessing; *ΔMAD* is the difference in *MAD* values between the target (T) and non-target (NT) subsets; ΔMAD_R is the ratio of *ΔMAD* to *MAD* for the non-target subset.

Table B.4. Comparison of Minimum and Maximum *ΔMAD* (Study 2B-SH)

Mean Theta Class	***J****	**Status**	**Mean Theta**	MAD_{NT}	MAD_{T}	ΔMAD	ΔMAD_{R}
Low	0	Max	-0.92	0.007	0.007	0.001	0.123
	6	Min	-0.34	0.058	0.020	0.038	1.937
	6	Max	-0.84	0.061	0.019	0.042	2.167
	9	Min	-0.35	0.083	0.028	0.056	2.012
	9	Max	-0.88	0.088	0.028	0.060	2.123
	12	Min	-0.35	0.108	0.039	0.069	1.746
Mid	0	Max	0.01	0.008	0.007	0.001	0.183
	6	Min	0.29	0.049	0.018	0.031	1.763
	6	Max	-0.24	0.057	0.019	0.038	1.974
	9	Min	0.32	0.070	0.024	0.046	1.893
	9	Max	-0.24	0.082	0.027	0.055	2.077
	12	Min	0.33	0.090	0.035	0.055	1.578
High	0	Max	0.56	0.008	0.007	0.002	0.226
	6	Min	0.99	0.033	0.014	0.019	1.418
	6	Max	0.33	0.048	0.018	0.030	1.710
	9	Min	1.00	0.051	0.018	0.033	1.768
	9	Max	0.40	0.068	0.023	0.045	1.948
	12	Min	0.98	0.064	0.028	0.036	1.315

Note. Max and Min status are, respectively, the maximum and minimum *ΔMAD* for the condition; *J** number of target items (Factor B); SH is spuriously high responding; *ΔMAD* is the difference in *MAD* values between the target (T) and non-target (NT) subsets; ΔMAD_R is the ratio of *ΔMAD* to *MAD* for the non-target subset.

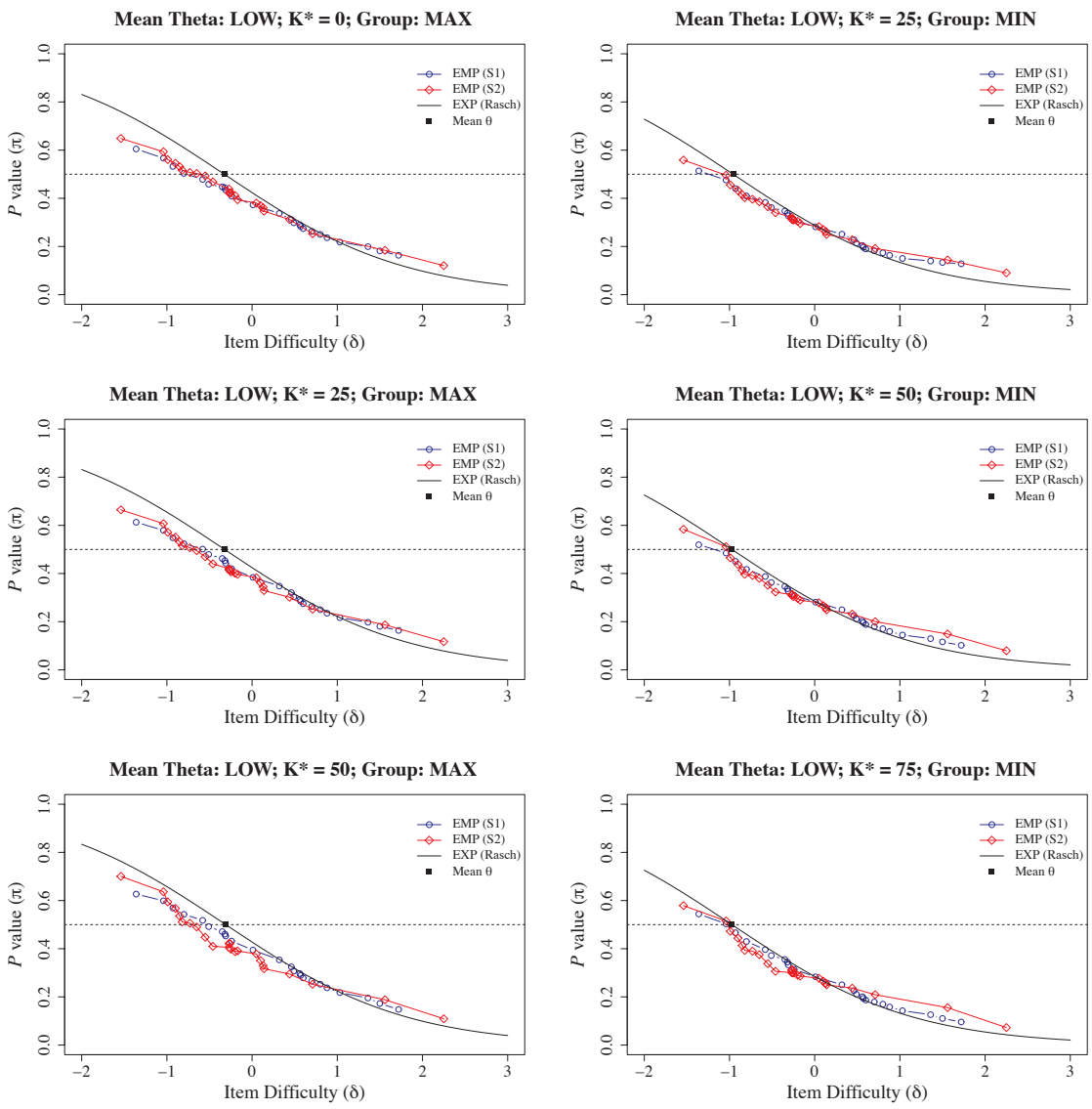


Figure B.1. Minimum and Maximum GRF Plots for LOW Mean Theta Class over Levels of Subgroup Homogeneity (*K**; Study 2A-GUESS)

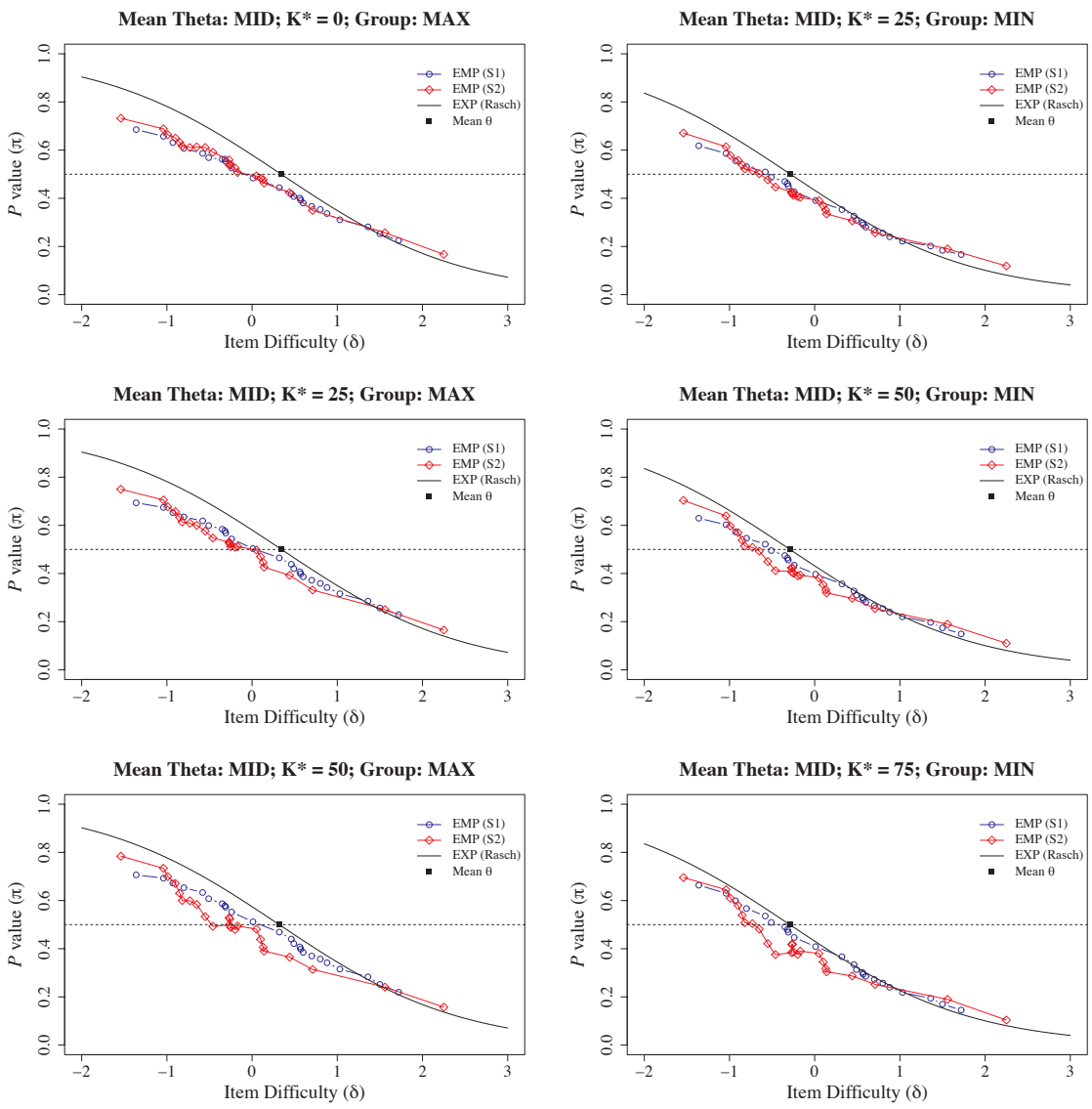


Figure B.2. Minimum and Maximum GRF Plots for MID Mean Theta Class over Levels of Subgroup Homogeneity (*K**; Study 2A-GUESS)

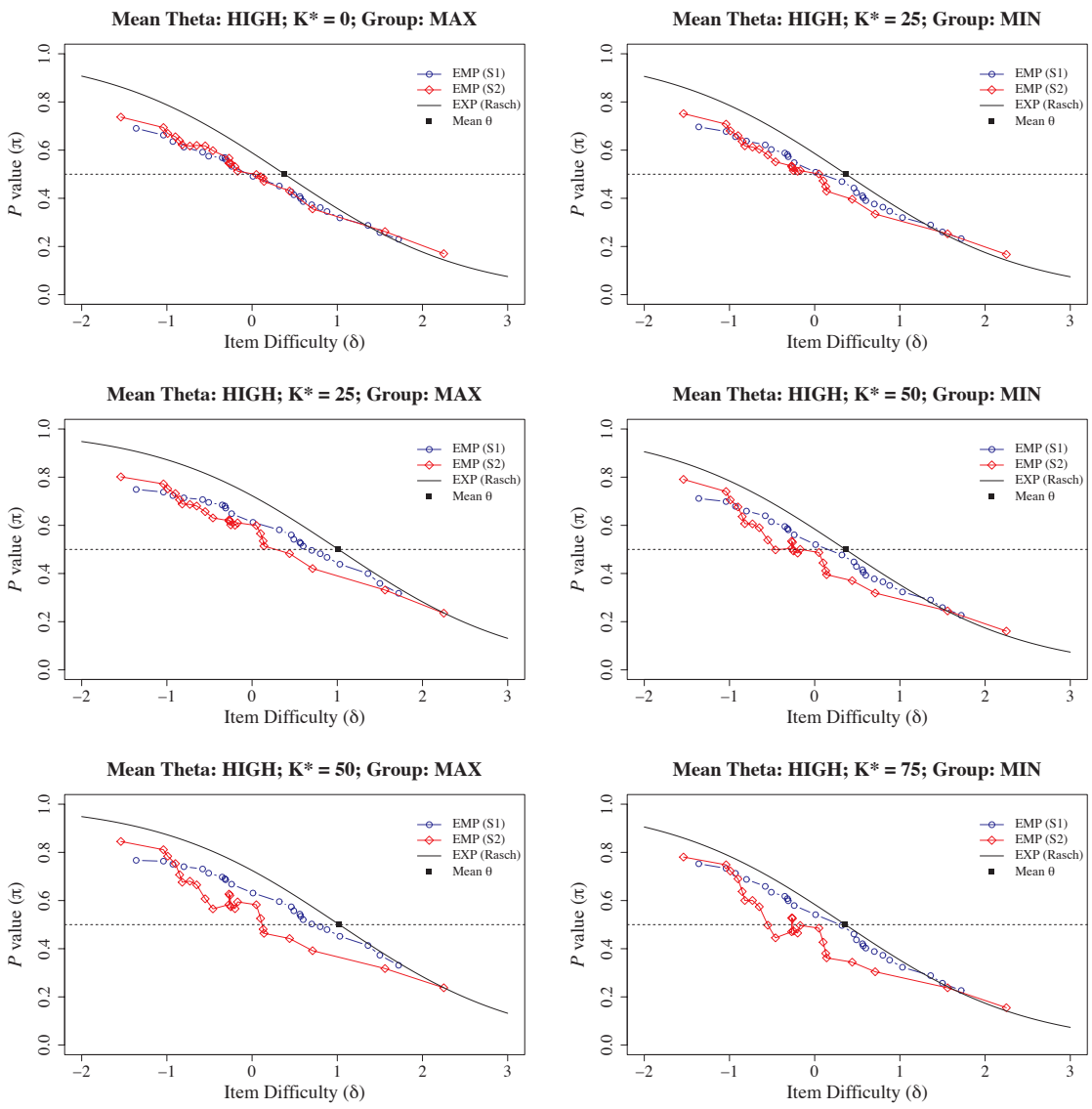


Figure B.3. Minimum and Maximum GRF Plots for HIGH Mean Theta Class over Levels of Subgroup Homogeneity (*K**; Study 2A-GUESS)

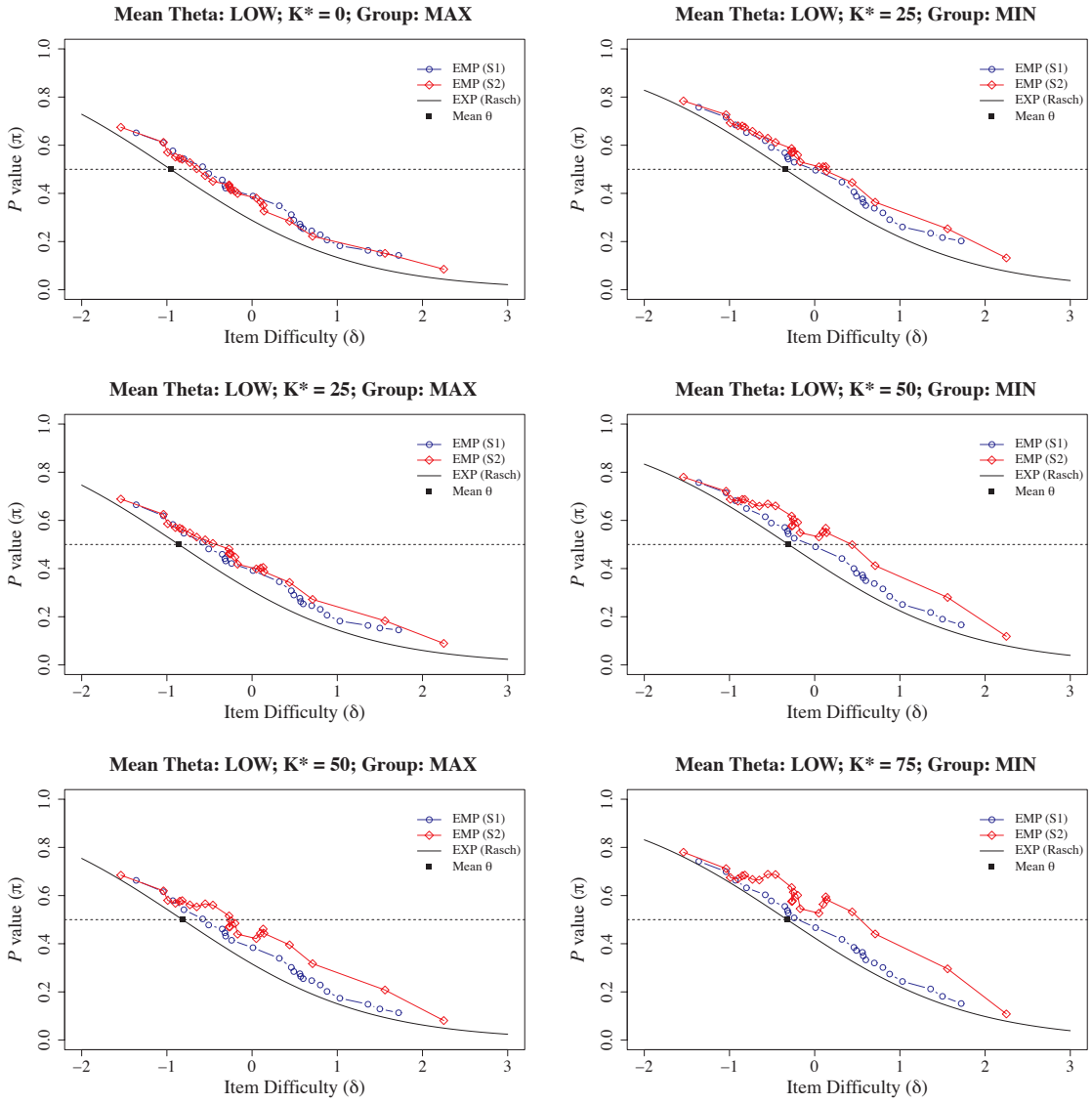


Figure B.4. Minimum and Maximum GRF Plots for LOW Mean Theta Class over Levels of Subgroup Homogeneity (*K**; Study 2A-SH)

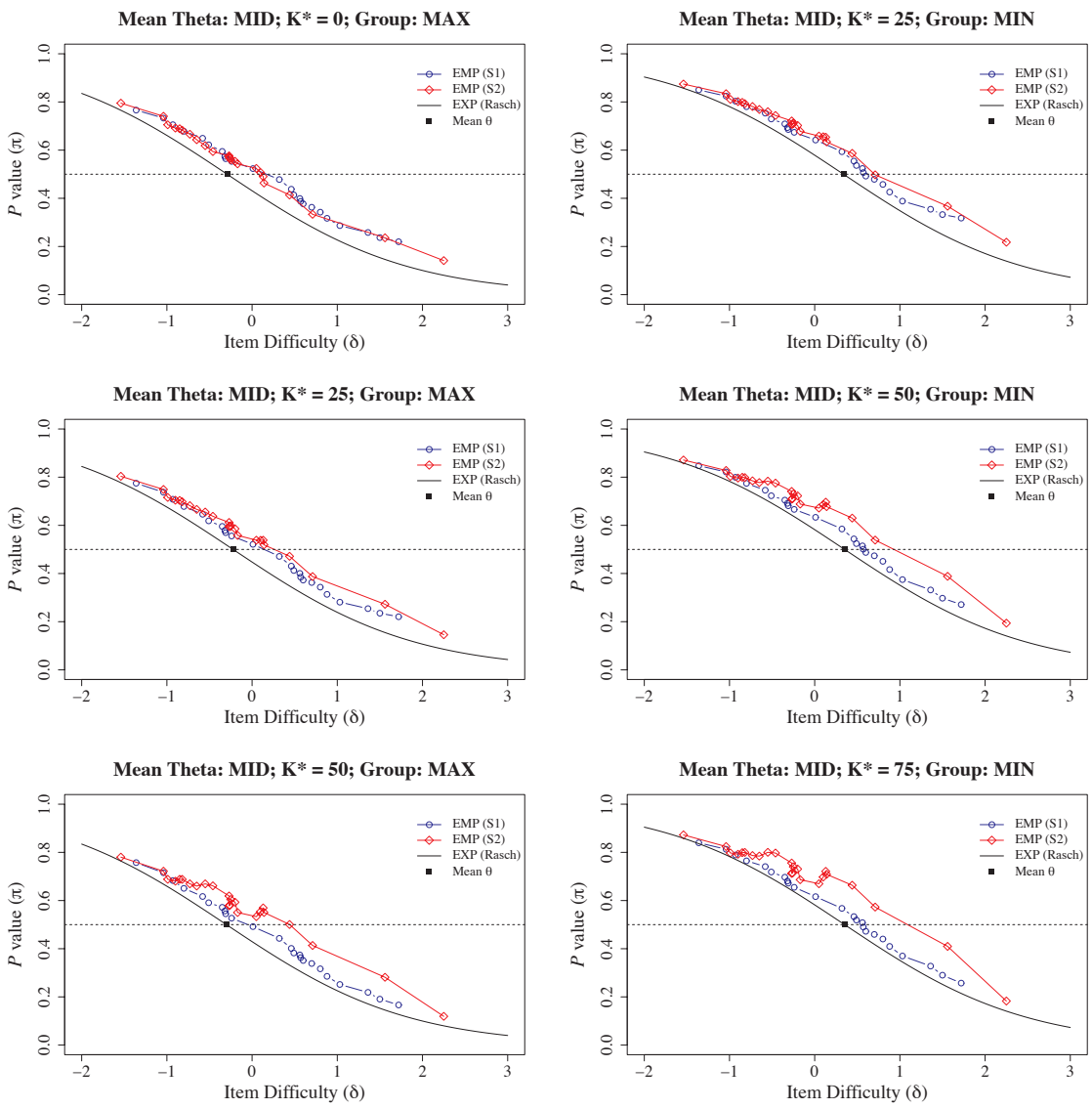


Figure B.5. Minimum and Maximum GRF Plots for MID Mean Theta Class over Levels of Subgroup Homogeneity (*K**; Study 2A-SH)

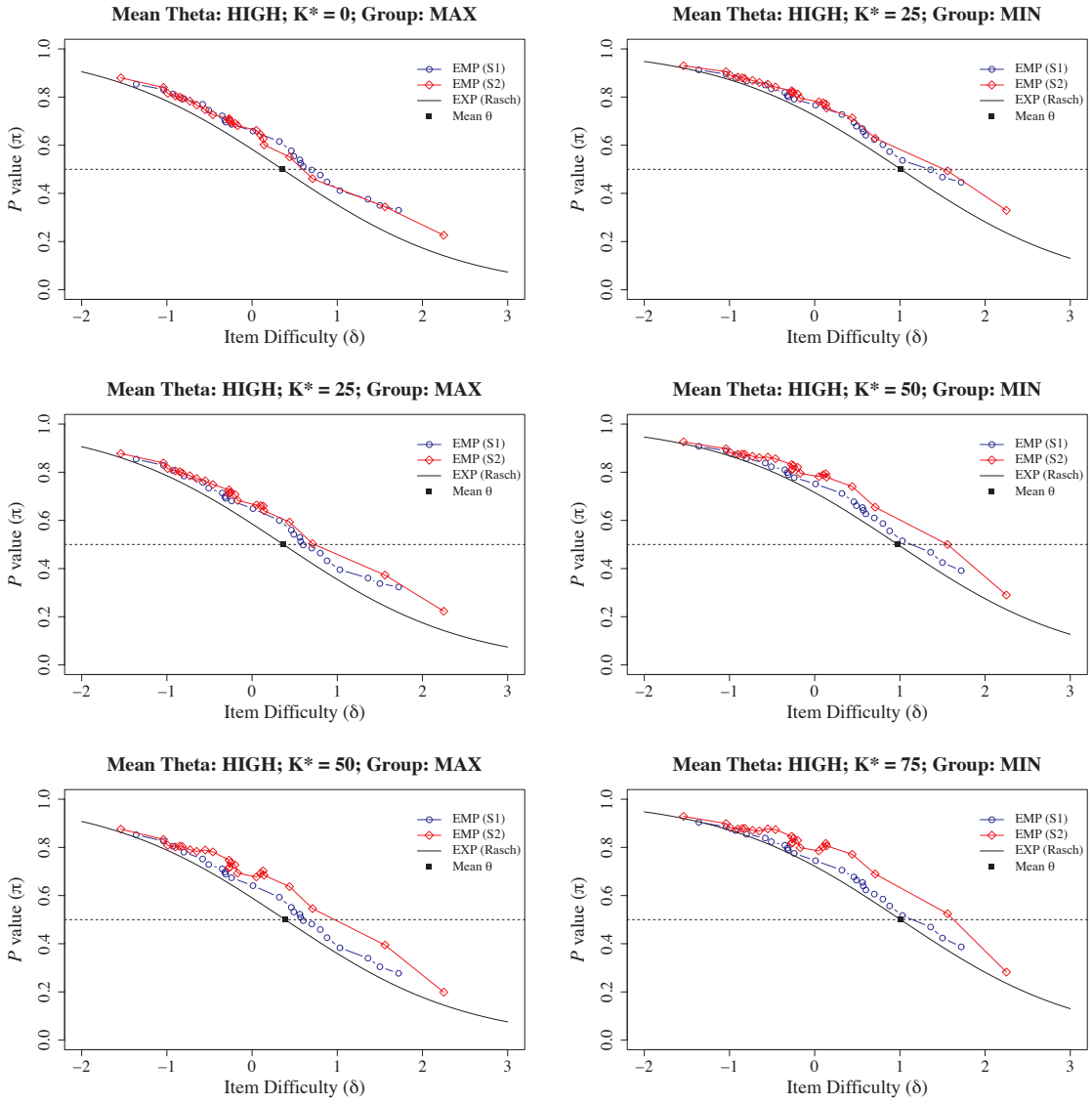


Figure B.6. Minimum and Maximum GRF Plots for HIGH Mean Theta Class over Levels of Subgroup Homogeneity (*K**; Study 2A-SH)

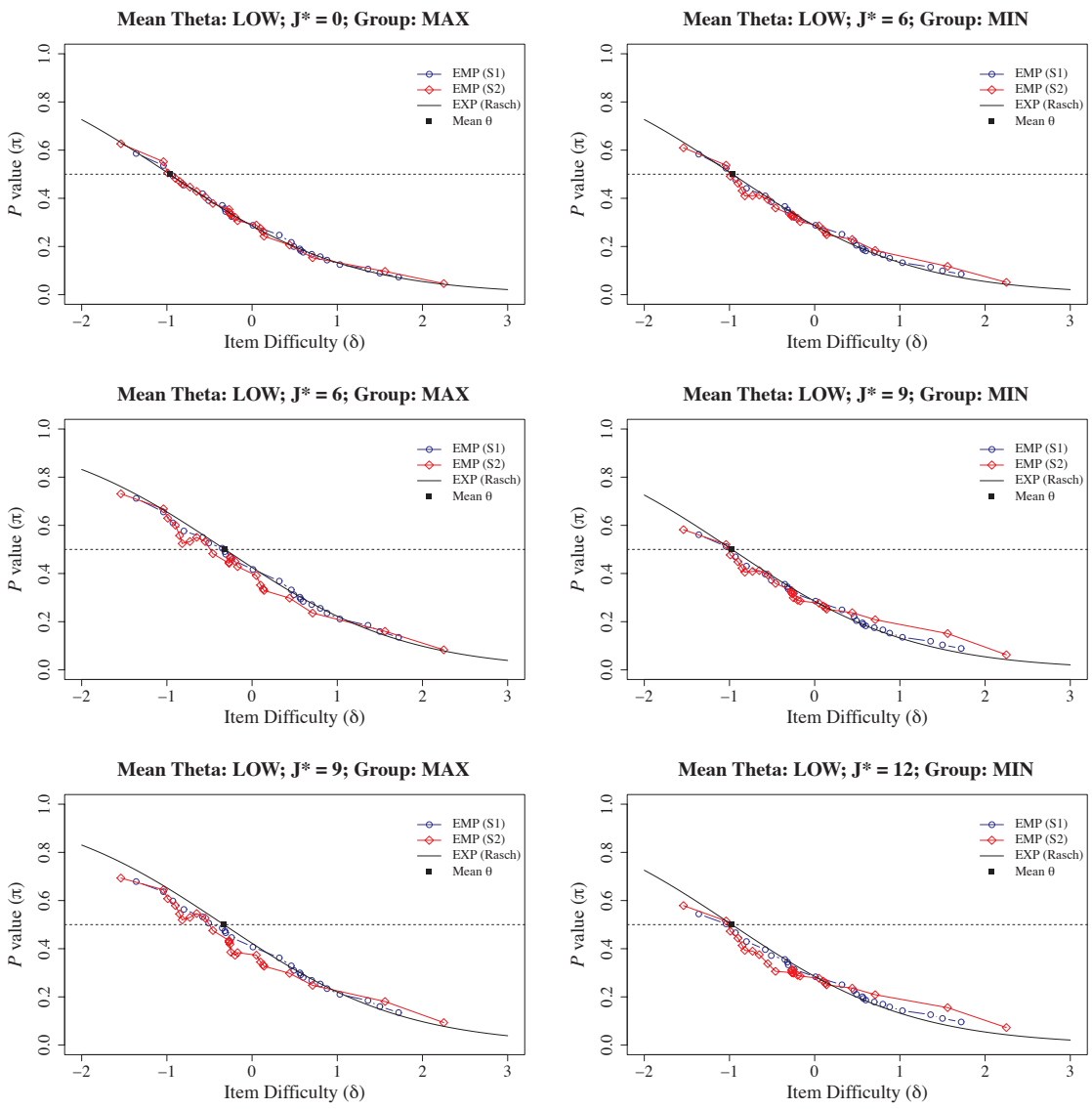


Figure B.7. Minimum and Maximum GRF Plots for LOW Mean Theta Class over Number of Target Items (*J**; Study 2B-GUESS)

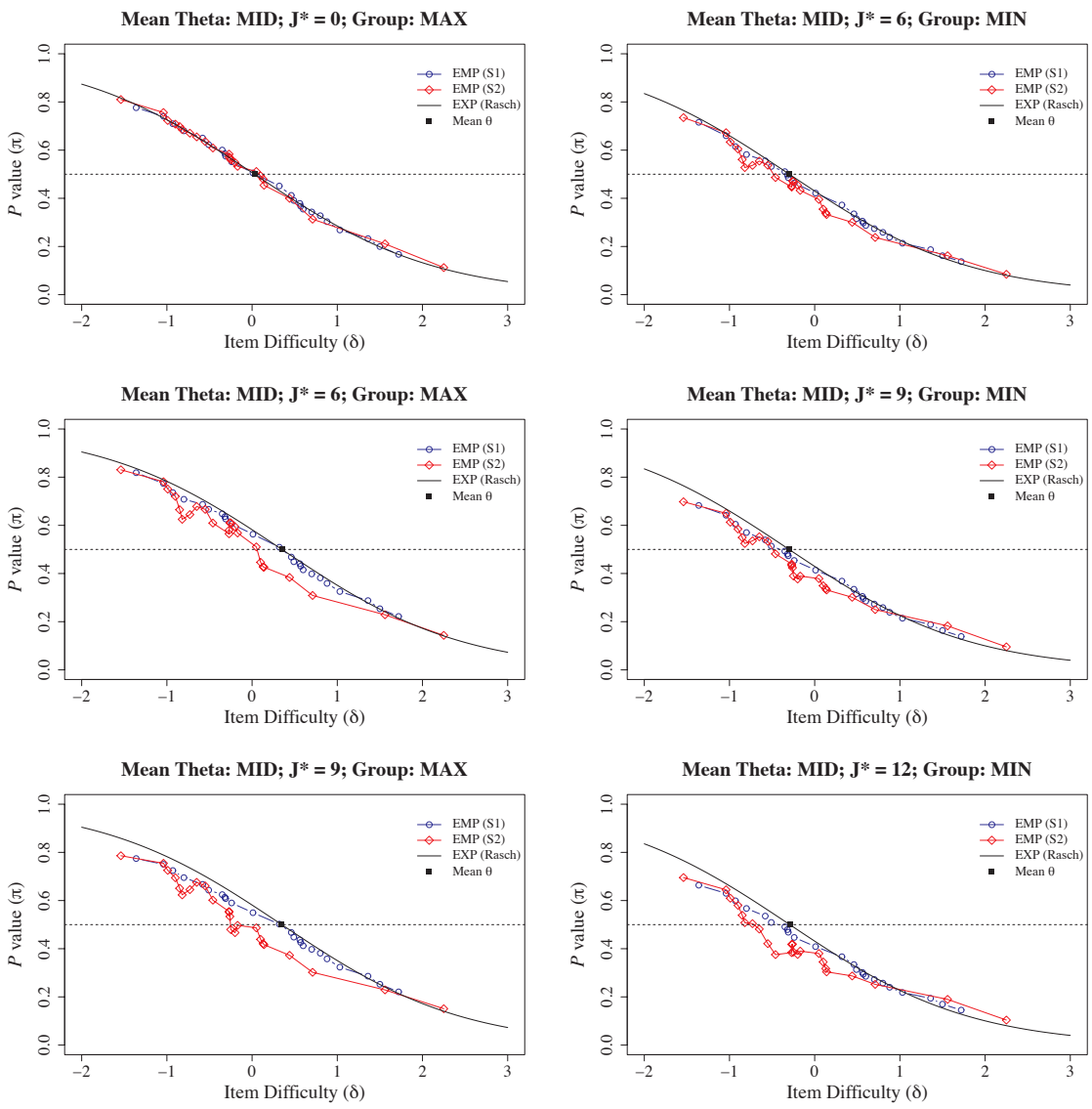


Figure B.8. Minimum and Maximum GRF Plots for MID Mean Theta Class over Number of Target Items (*J**; Study 2B-GUESS)

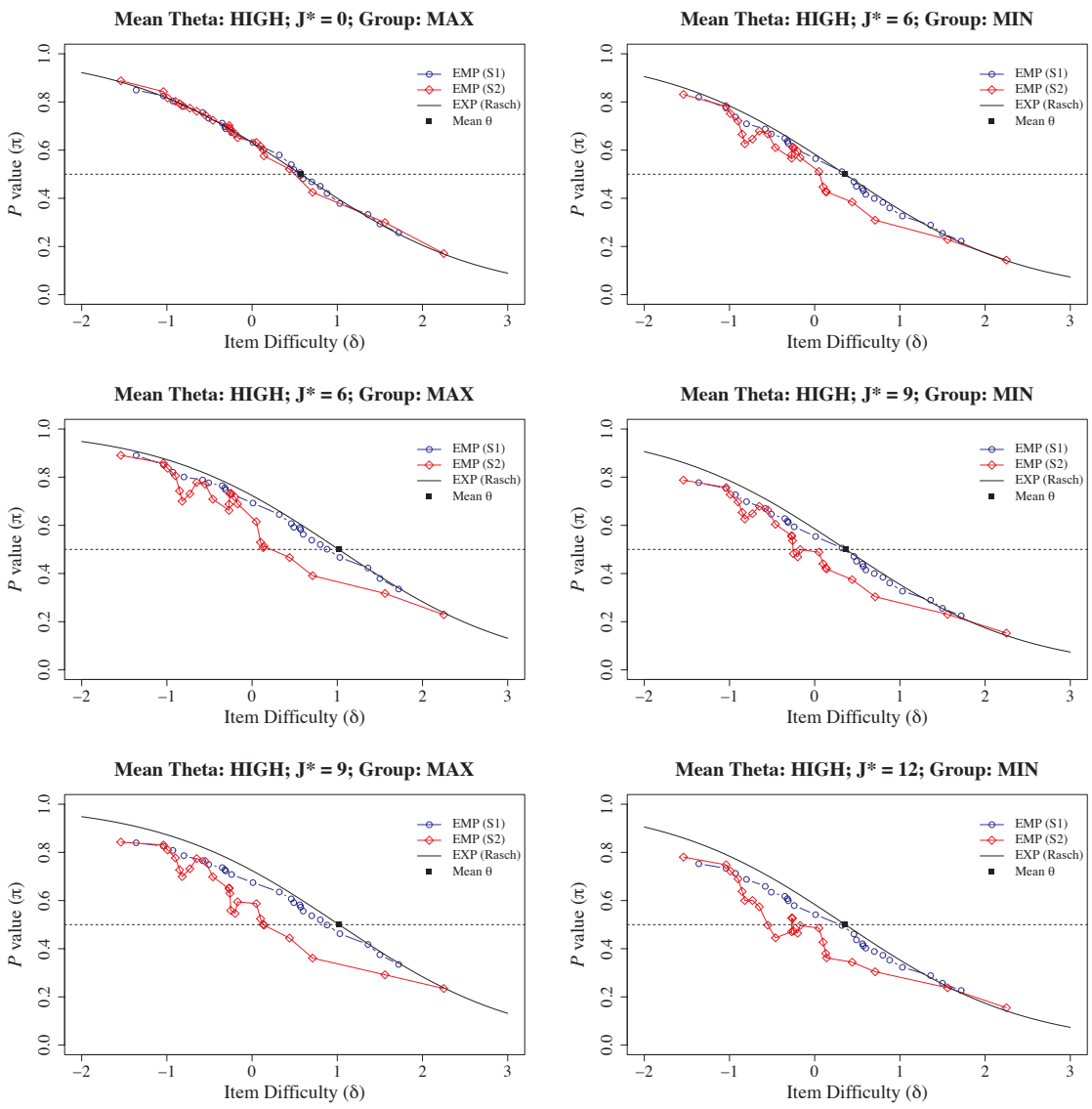


Figure B.9. Minimum and Maximum GRF Plots for HIGH Mean Theta Class over Number of Target Items (*J**; Study 2B-GUESS)

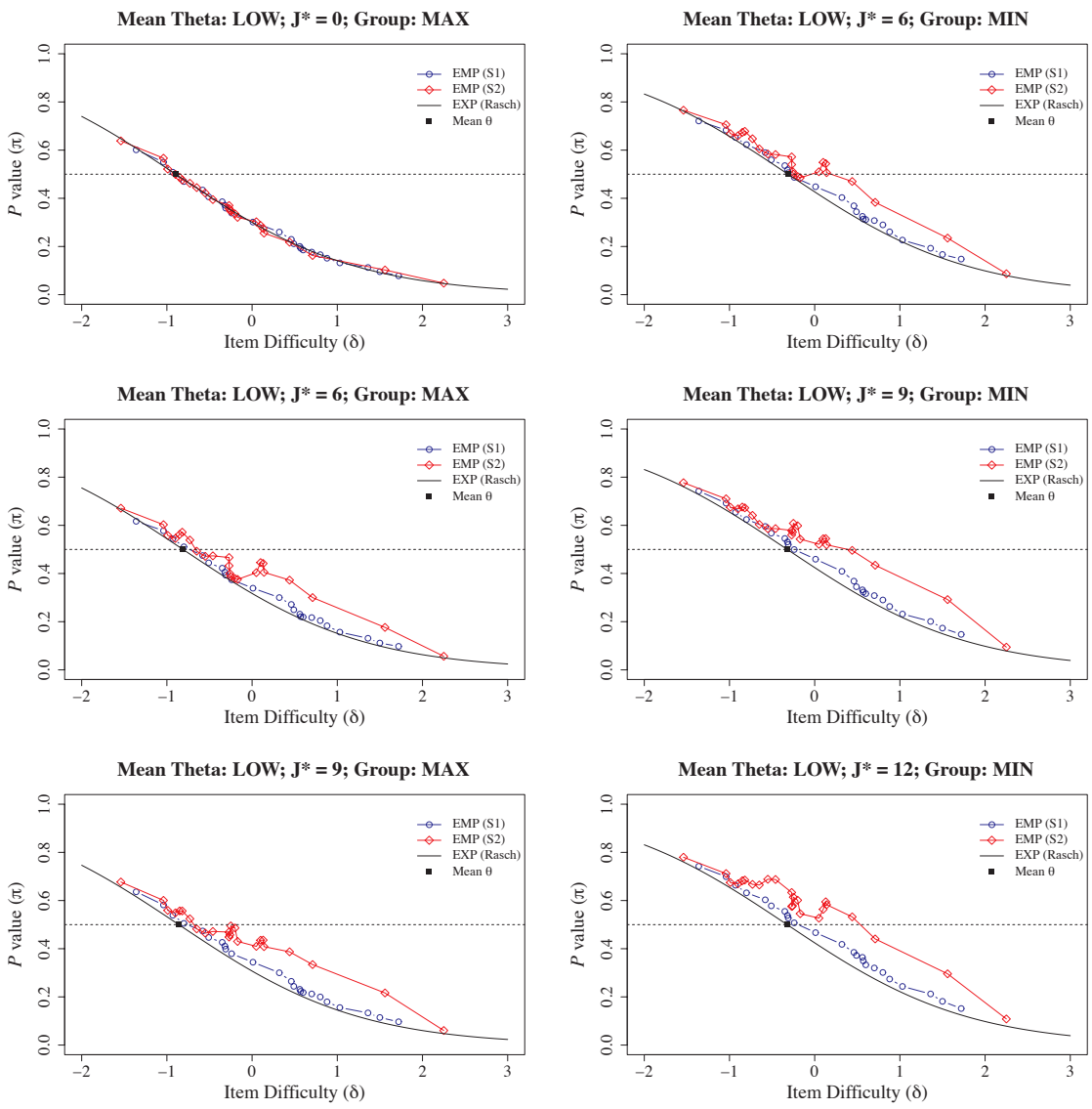


Figure B.10. Minimum and Maximum GRF Plots for LOW Mean Theta Class over Number of Target Items (*J**; Study 2B-SH)

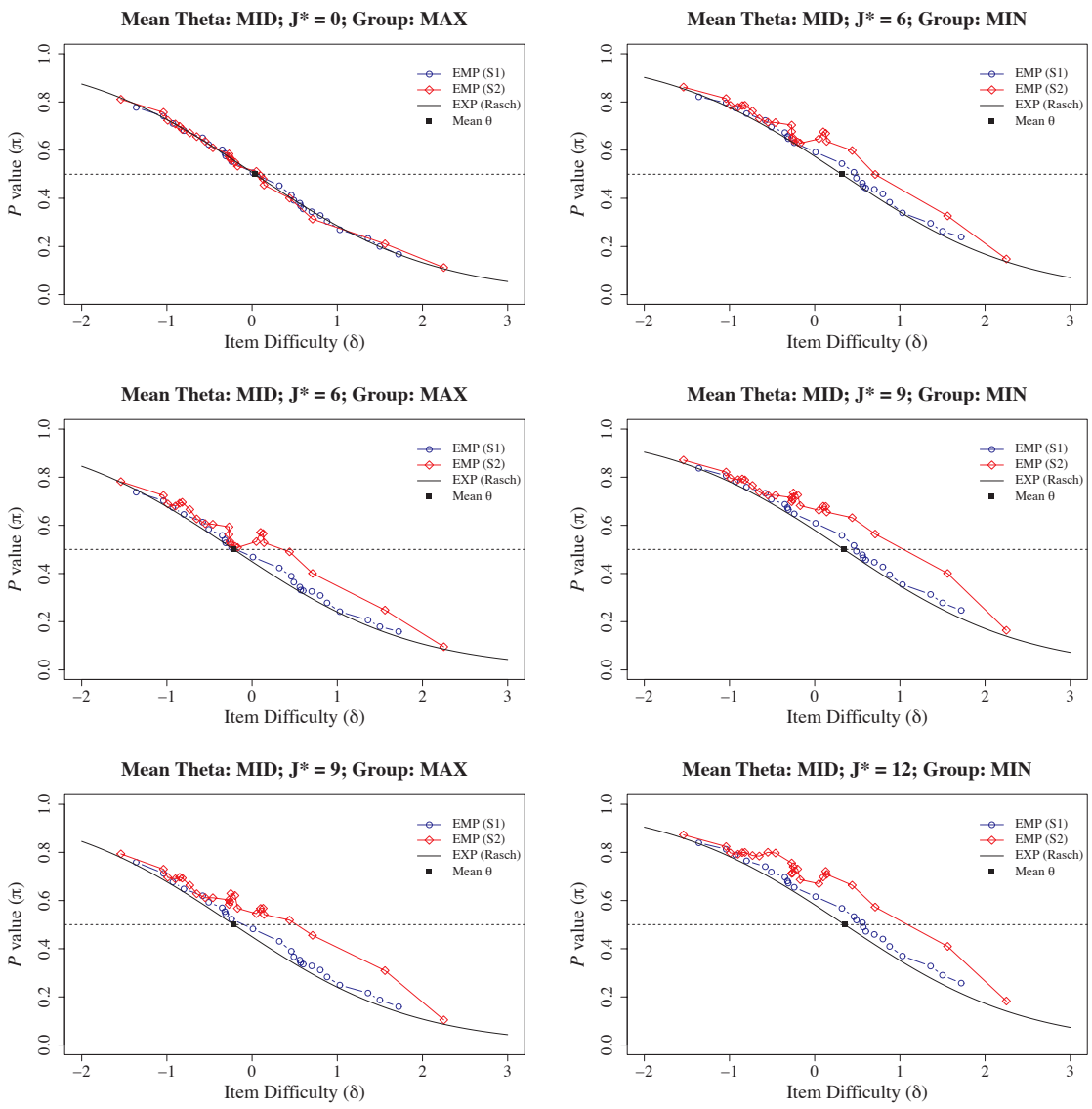


Figure B.11. Minimum and Maximum GRF Plots for MID Mean Theta Class over Number of Target Items (*J**; Study 2B-SH)

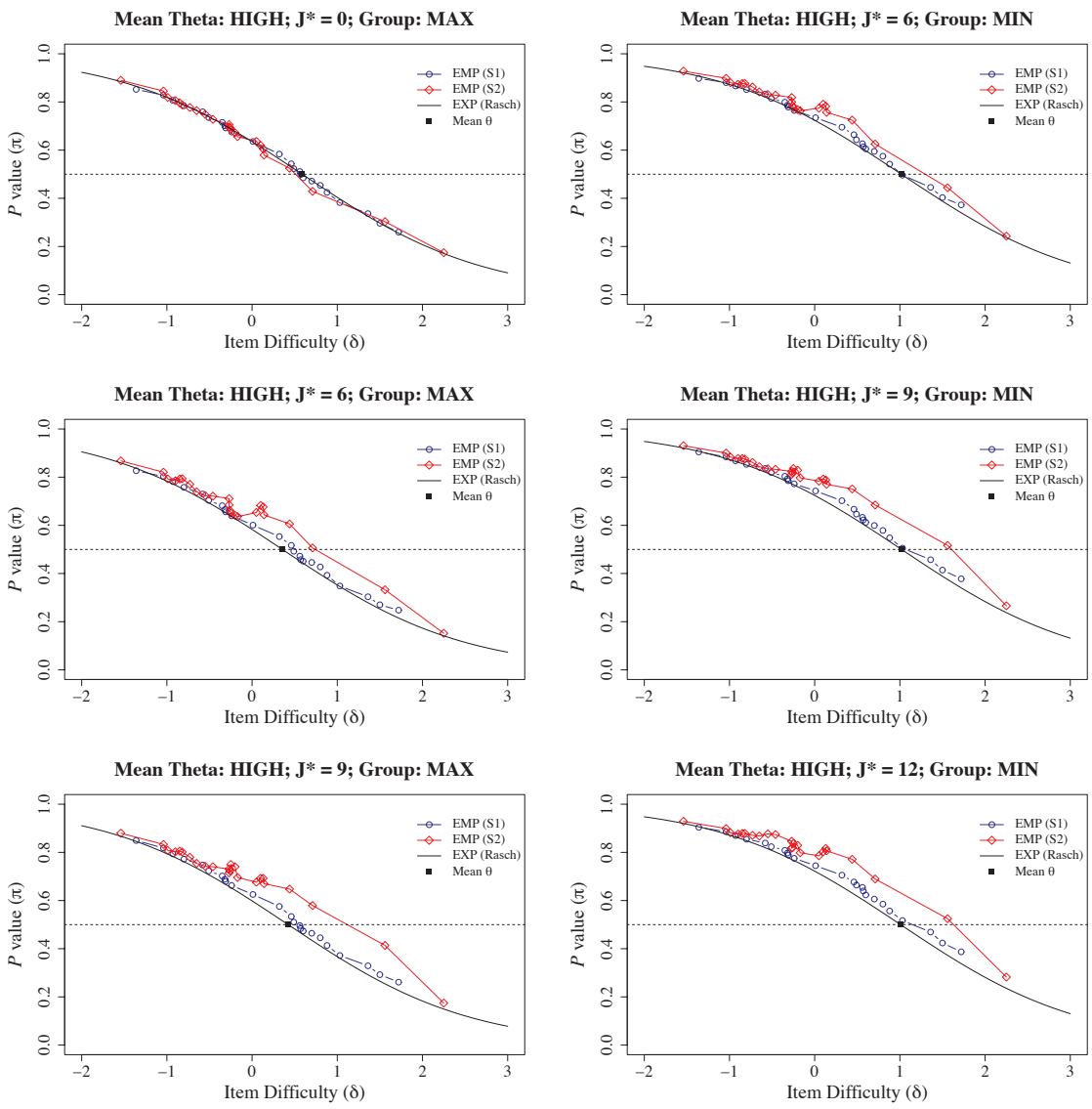


Figure B.12. Minimum and Maximum GRF Plots for HIGH Mean Theta Class over Number of Target Items (*J**; Study 2B-SH)

APPENDIX C

SUMMARY TABLES FOR STUDY 2 OUTCOME VARIABLES

Table C.1. Mean *MAD* Outcomes by Condition for Sub-Study 2A-GUESS

Mean Theta Class	*K**	*N* Subgoups	MAD_{Target}		$MAD_{NonTarget}$		ΔMAD		ΔMAD_R	
			Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Low	0	53	0.045	0.009	0.037	0.004	0.007	0.005	0.183	0.103
	25	49	0.049	0.010	0.032	0.004	0.017	0.006	0.511	0.136
	50	61	0.051	0.011	0.026	0.003	0.024	0.007	0.907	0.180
	75	52	0.058	0.014	0.020	0.002	0.039	0.011	1.926	0.355
Mid	0	51	0.076	0.009	0.061	0.009	0.015	0.001	0.255	0.027
	25	50	0.085	0.011	0.051	0.007	0.033	0.004	0.656	0.017
	50	42	0.091	0.011	0.043	0.006	0.048	0.006	1.121	0.017
	75	60	0.103	0.013	0.031	0.004	0.072	0.009	2.344	0.049
High	0	46	0.103	0.007	0.089	0.008	0.014	0.002	0.162	0.033
	25	51	0.121	0.011	0.076	0.008	0.045	0.002	0.589	0.032
	50	47	0.134	0.013	0.065	0.007	0.069	0.006	1.071	0.014
	75	38	0.146	0.013	0.046	0.005	0.100	0.008	2.180	0.040

Note. *K** is group homogeneity (Factor A); GUESS is simulated guessing; ΔMAD is the difference in *MAD* values between subset GRFs (target and non-target item subsets); ΔMAD_R is the ratio of ΔMAD to *MAD* for the non-target subset.

Table C.2. Mean *MAD* Outcomes by Condition for Sub-Study 2A-SH

Mean Theta Class	*K**	*N* Subgoups	MAD_{Target}		$MAD_{NonTarget}$		ΔMAD		ΔMAD_R	
			Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Low	0	45	0.072	0.001	0.069	0.074	-0.007	0.001	-0.086	0.013
	25	39	0.086	0.001	0.083	0.088	0.021	0.001	0.323	0.008
	50	57	0.101	0.002	0.098	0.103	0.050	0.001	0.959	0.016
	75	47	0.113	0.002	0.108	0.116	0.073	0.002	1.836	0.040
Mid	0	45	0.064	0.003	0.058	0.069	-0.011	0.001	-0.143	0.016
	25	59	0.076	0.004	0.069	0.082	0.015	0.002	0.248	0.026
	50	38	0.091	0.004	0.081	0.097	0.043	0.003	0.879	0.043
	75	55	0.099	0.006	0.090	0.108	0.061	0.004	1.640	0.072
High	0	60	0.049	0.005	0.041	0.057	-0.013	0.000	-0.217	0.018
	25	52	0.060	0.006	0.050	0.069	0.008	0.002	0.153	0.026
	50	55	0.070	0.006	0.059	0.079	0.030	0.003	0.747	0.041
	75	48	0.078	0.007	0.064	0.090	0.047	0.005	1.459	0.075

Note. *K** is group homogeneity (Factor A); SH is spuriously high responding; *ΔMAD* is the difference in *MAD* values between subset GRFs (target and non-target item subsets); ΔMAD_R is the ratio of *ΔMAD* to *MAD* for the non-target subset.

Table C.3. Mean *MAD* Outcomes by Condition for Sub-Study 2B-GUESS

Mean Theta Class	*J**	*N* Subgoups	MAD_{Target}		$MAD_{NonTarget}$		ΔMAD		ΔMAD_R	
			Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Low	0	39	0.007	0.000	0.007	0.000	0.000	0.000	0.068	0.025
	6	42	0.027	0.006	0.010	0.001	0.017	0.005	1.633	0.312
	9	52	0.038	0.008	0.015	0.002	0.023	0.006	1.530	0.212
	12	52	0.058	0.014	0.020	0.002	0.039	0.011	1.926	0.355
Mid	0	47	0.008	0.000	0.007	0.000	0.001	0.000	0.140	0.024
	6	46	0.050	0.008	0.015	0.002	0.035	0.006	2.388	0.164
	9	51	0.074	0.010	0.023	0.003	0.051	0.008	2.179	0.104
	12	60	0.103	0.013	0.031	0.004	0.072	0.009	2.344	0.049
High	0	64	0.008	0.000	0.007	0.000	0.001	0.000	0.178	0.028
	6	62	0.075	0.007	0.022	0.002	0.054	0.005	2.484	0.041
	9	47	0.110	0.011	0.034	0.003	0.076	0.007	2.204	0.021
	12	38	0.146	0.013	0.046	0.005	0.100	0.008	2.180	0.040

Note. *J** is group homogeneity (Factor B); GUESS is simulated guessing; ΔMAD is the difference in *MAD* values between subset GRFs (target and non-target item subsets); ΔMAD_R is the ratio of ΔMAD to *MAD* for the non-target subset.

Table C.4. Mean *MAD* Outcomes by Condition for Sub-Study 2B-SH

Mean Theta Class	*J**	*N* Subgoups	MAD_{Target}		$MAD_{NonTarget}$		ΔMAD		ΔMAD_R	
			Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Low	0	52	0.007	0.000	0.007	0.000	0.000	0.000	0.065	0.025
	6	57	0.060	0.001	0.019	0.000	0.041	0.001	2.115	0.054
	9	53	0.086	0.001	0.028	0.000	0.058	0.001	2.085	0.032
	12	47	0.113	0.002	0.040	0.000	0.073	0.002	1.836	0.040
Mid	0	45	0.008	0.000	0.007	0.000	0.001	0.000	0.140	0.029
	6	42	0.054	0.002	0.019	0.001	0.035	0.002	1.819	0.076
	9	53	0.076	0.003	0.026	0.001	0.050	0.002	1.935	0.044
	12	55	0.099	0.006	0.037	0.001	0.061	0.004	1.640	0.072
High	0	53	0.008	0.000	0.007	0.000	0.001	0.000	0.164	0.035
	6	51	0.042	0.004	0.016	0.001	0.026	0.003	1.600	0.101
	9	44	0.060	0.006	0.021	0.002	0.039	0.004	1.845	0.057
	12	48	0.078	0.007	0.032	0.002	0.047	0.005	1.459	0.075

Note. *J** is group homogeneity (Factor B); SH is spuriously high responding; ΔMAD is the difference in *MAD* values between subset GRFs (target and non-target item subsets); ΔMAD_R is the ratio of ΔMAD to *MAD* for the non-target subset.

Table C.5. Mean Mean Person Fit Statistics by Sub-Study and Condition (Study 2)

Sub-Study (Factor)	Mean Theta Class	Factor Level	*N* Subgroups	U_M		W_M		UB_M	
				Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***
2A-GUESS	Low	0	53	1.115	0.015	1.058	0.004	2.444	0.138
(*K**)		25	49	1.112	0.015	1.057	0.004	2.499	0.146
		50	61	1.100	0.014	1.051	0.004	2.512	0.162
		75	52	1.093	0.015	1.049	0.005	2.589	0.206
	Mid	0	51	1.152	0.037	1.114	0.031	3.274	0.373
		25	50	1.150	0.038	1.112	0.032	3.428	0.447
		50	42	1.134	0.031	1.102	0.028	3.511	0.421
		75	60	1.132	0.032	1.102	0.029	3.714	0.487
	High	0	46	1.348	0.080	1.255	0.052	4.922	0.598
		25	51	1.376	0.105	1.275	0.070	5.568	0.897
		50	47	1.360	0.105	1.273	0.073	5.966	1.036
		75	38	1.319	0.085	1.249	0.061	6.037	0.960
2A-SH	Low	0	45	1.169	0.039	1.114	0.028	3.595	0.130
(*K**)		25	39	1.180	0.040	1.123	0.029	3.768	0.139
		50	57	1.170	0.046	1.123	0.035	3.949	0.177
		75	47	1.167	0.043	1.122	0.033	4.035	0.169
	Mid	0	45	1.053	0.031	1.031	0.022	3.150	0.132
		25	59	1.049	0.030	1.027	0.022	3.273	0.128
		50	38	1.051	0.029	1.030	0.023	3.438	0.136
		75	55	1.041	0.034	1.023	0.027	3.463	0.168

Table C.5 (continued).

	High	0	60	0.958	0.022	0.962	0.016	2.697	0.111
		25	52	0.961	0.022	0.963	0.017	2.819	0.125
		50	55	0.956	0.020	0.956	0.015	2.911	0.123
		75	48	0.958	0.021	0.957	0.016	2.987	0.146
2B-GUESS	Low	0	39	1.002	0.001	1.001	0.001	2.012	0.008
(*J**)		6	42	1.032	0.006	1.019	0.002	2.127	0.039
		9	52	1.069	0.012	1.035	0.003	2.263	0.085
		12	52	1.093	0.015	1.049	0.005	2.589	0.206
	Mid	0	47	1.002	0.001	1.001	0.001	2.009	0.012
		6	46	1.051	0.017	1.042	0.015	2.403	0.143
		9	51	1.092	0.023	1.071	0.020	2.877	0.263
		12	60	1.132	0.032	1.102	0.029	3.714	0.487
	High	0	64	1.000	0.001	1.000	0.000	2.028	0.016
		6	62	1.142	0.039	1.116	0.030	3.087	0.272
		9	47	1.242	0.067	1.187	0.048	4.312	0.596
		12	38	1.319	0.085	1.249	0.061	6.037	0.960
2B-SH	Low	0	52	1.002	0.001	1.001	0.001	2.012	0.008
(*J**)		6	57	1.084	0.023	1.068	0.019	2.590	0.067
		9	53	1.132	0.031	1.099	0.023	3.226	0.092
		12	47	1.167	0.043	1.122	0.033	4.035	0.169
	Mid	0	45	1.002	0.001	1.001	0.001	2.009	0.012
		6	42	1.020	0.014	1.014	0.013	2.394	0.049
		9	53	1.029	0.020	1.018	0.016	2.871	0.084
		12	55	1.041	0.034	1.023	0.027	3.463	0.168

Table C.5 (continued).

High	0	53	1.000	0.001	1.000	0.000	2.023	0.017
	6	51	0.976	0.010	0.976	0.009	2.240	0.039
	9	44	0.965	0.016	0.966	0.014	2.572	0.085
	12	48	0.958	0.021	0.957	0.016	2.987	0.146

Note. K^* is subgroup homogeneity (Factor A); J^* is number of target items (Factor B); GUESS is simulated guessing; SH is simulated spuriously high responding; U_M is the mean outfit statistic for the subgroup, W_M is the mean infit statistic for the subgroup, UB_M is mean between-fit statistic for the subgroup.

APPENDIX D

THETA AND FIT STATISTIC DENSITY PLOTS (STUDY 3)

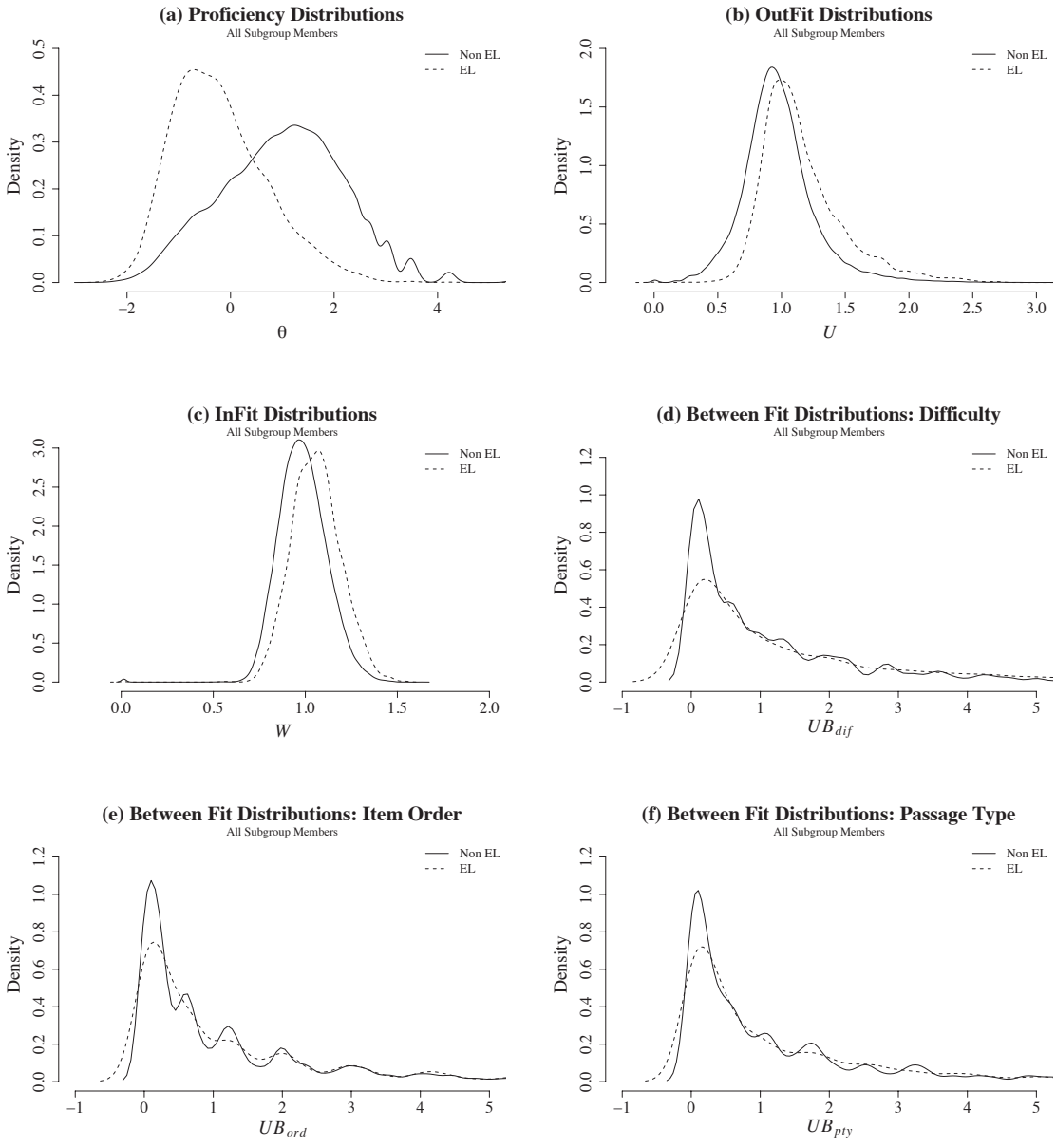


Figure D.1. Proficiency and Person Fit Distribution Comparison between EL and Non-EL Examinees

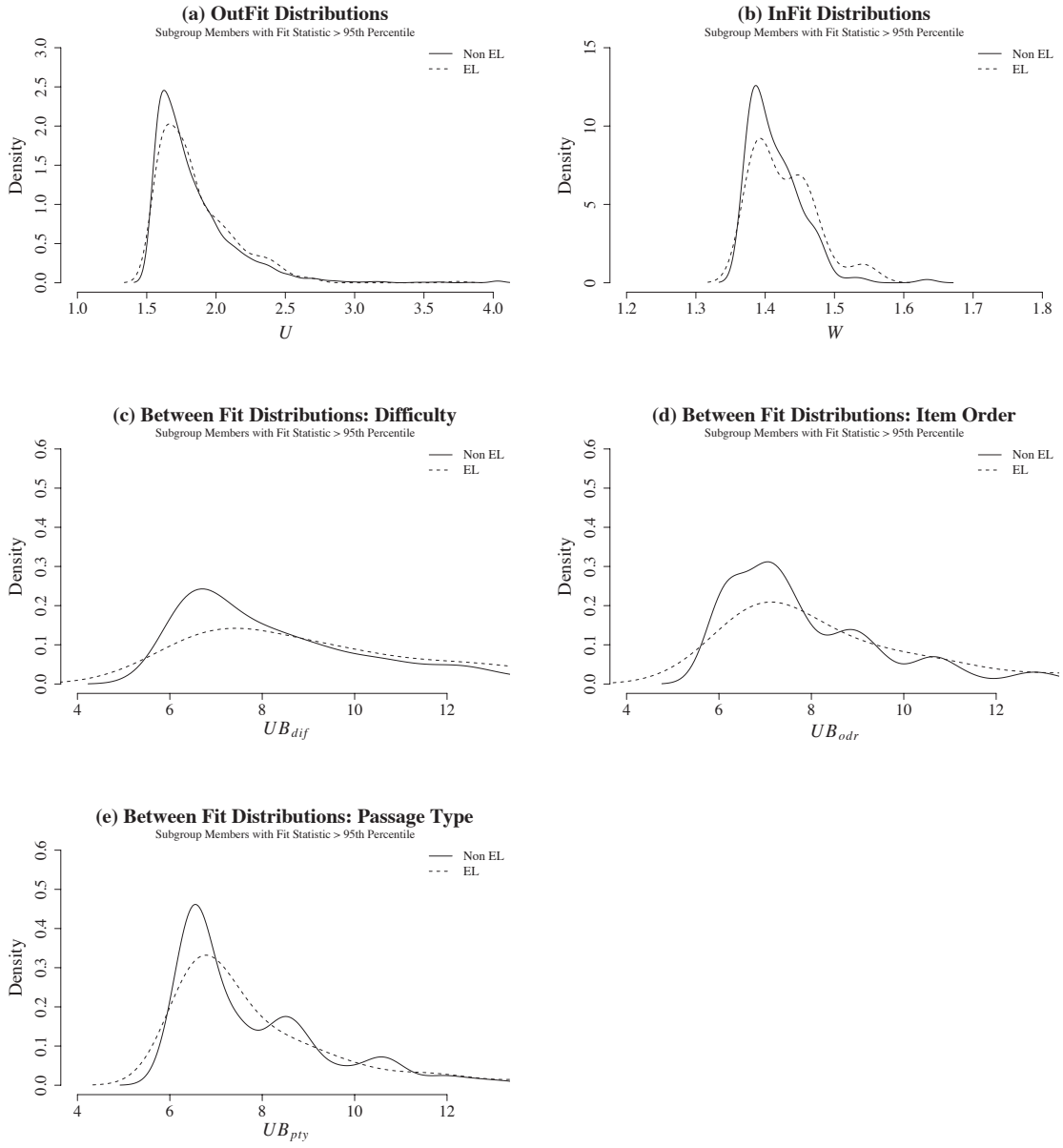


Figure D.2. Person Fit Distribution Comparison between Aberrant EL and Aberrant Non-EL Examinees